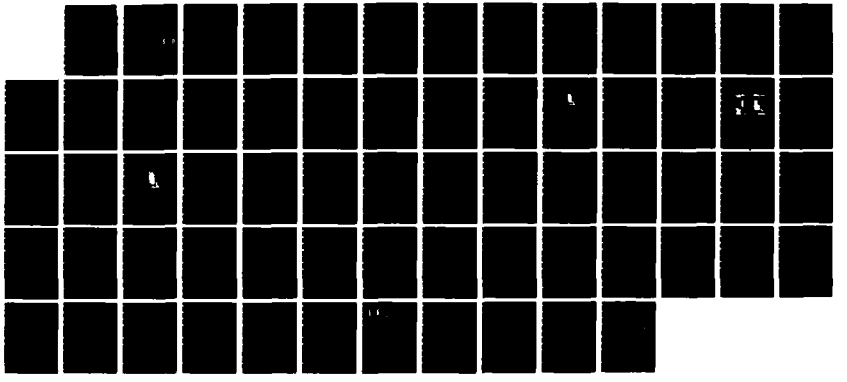
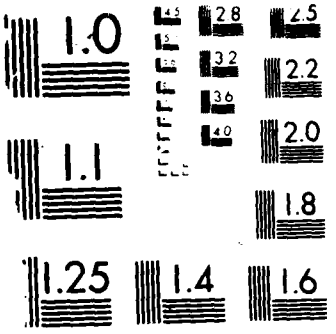
AD-A180 247 SPEECH RECOGNITION: ACOUSTIC PHONETIC AND LEXICAL(U) 1/1
MASSACHUSETTS INST OF TECH CAMBRIDGE RESEARCH LAB OF
ELECTRONICS V W ZUE 15 SEP 86 N00014-82-K-0727

UNCLASSIFIED NL





ANNUAL PROGRESS REPORT

SPEECH RECOGNITION:
Acoustic, Phonetic and Lexical

Office of Naval Research
Contract N00014-82-K-0727

Covering the Period
1 July 1985 – 30 June 1986

Submitted by:

Victor W. Zue

September 15, 1986

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Research Laboratory of Electronics
Cambridge, Massachusetts 02139

RESEARCH LABORATORY OF ELECTRONICS, 36-591
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

September 15, 1986

Director
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22200

Attention: Program Management

This letter is the Annual Progress Report for our research program supported under DARPA-ONR Contract N00014-82-K-0727.

During the period of 1 July 1985 to 30 June 1986, we have continued to make progress on the acquisition of acoustic-phonetic and lexical knowledge. Specifically:

- We have concluded our studies of lexical stress and improved the performamce of the lexical stress recognition system. The system is composed of two parts: a syllable detector and a stress determiner. A number of modifications were made to the syllable detector, including the introduction of more robust intervocalic consonant detectors, new algorithms for sonorant detection, and improvements in code to shorten run times and increase user flexibility for system development. The system now runs approximately three times faster, detects sonorants more accurately, makes fewer false insertions, and is more flexible.

- We have conducted experiments to quantify the influence of phonetic context, including syllable structure, on the acoustic properties of stop consonants. Our results indicate that both syllable structure and phonemic context play a significant role in determining whether a stop will be released, unreleased, or deleted altogether. By continuing to study such contextual variations and their acoustic consquences, we hope to eventually implement a computational framework that incorporates context knowledge in phonemic decoding.

- We have undertaken an investigation to capture the knowledge that humans use to read spectrograms, and to apply this knowledge to the creation of an expert system. Humans are able to read spectrograms by extracting and then integrating the relevant acoustic features, using rules that relate the underlying phonetic forms to their acoustic manifestations. To test the feasibility of

developing a computer system that mimics such a process, we selected a task of identifying stop consonants drawn from continuous speech. Our preliminary results indicate that machine performance comparable to that of human experts can be attained.

- We have begun development of a system that applies vision techniques to extract acoustic patterns in speech spectrograms. By processing a spectrographic image through a set of edge detectors and combining their outputs, the system obtains two-dimensional objects that characterize the formant patterns and general spectral properties of vowels and consonants. Preliminary evidence suggests that the visual characterizations produced by this processing technique may provide an effective alternative to traditional descriptions of acoustic-phonetic events.

- We have initiated development of an articulatory synthesizer, LAMINAR, capable of synthesizing speech from different vocal tract configurations. This new speech research tool takes an articulatory configuration in the form of an acoustic tube, and generates the resulting acoustic output. With continued development, the system could realistically model many time-varying articulatory gestures, thus providing a useful mechanism for speech production experiments.

We are including with this report copies of the following publications, in the form of theses and papers presented at various conferences, written with ONR support during this contracting period:

- Chen, F. R., "Lexical Access and Verification in a Broad Phonetic Approach to Continuous Digit Recognition."

- Huttenlocher, D. P., "A Broad Phonetic Classifier."

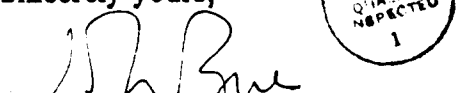
- Leung, H. C., and V. W. Zue, "Visual Characterization of Speech Spectrograms."

- Unverferth, J. E., "Improvements to and Extensions of an Automatic Lexical Stress Determiner."

- Zue, V. W., "Utilizing Speech-Specific Knowledge in Automatic Speech Recognition."

- Zue, V. W., and L. F. Lamel, "An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition."

Sincerely yours,



Victor W. Zue
Principal Investigator

Enc.

Improvements to and Extensions of an Automatic Lexical Stress Determiner

by

John Edward Unverferth III

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

BACHELOR OF SCIENCE
IN ELECTRICAL SCIENCE AND ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 1986

Signature of Author: _____
Department of Electrical Engineering and Computer Science
May 8, 1986

Certified by: _____
Victor W. Zue
Faculty Thesis Supervisor

Accepted by: _____
David Adler
Chairman, Department Committee

Improvements to and Extensions of an Automatic Lexical Stress Determiner

by

John Edward Unverferth III

Submitted to the Department of Electrical Engineering and Computer Science on May 9, 1986, in partial fulfillment of the requirements for the degree of Bachelor of Science

Abstract

As part of her Master's Thesis, Aull constructed a Lexical Stress Determiner for discrete words. Her system was designed to determine the number of syllables and the lexical stress pattern in discrete words. The purpose of this thesis is to make her system more robust, both from a programmer's point of view as well as from a performance and reliability perspective.

Thesis Supervisor: Dr. Victor W. Zue

Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank Victor Zue, my thesis advisor who gave me a great deal of help and encouragement, not only on my thesis but throughout the last two years. I would also like to especially thank Scott Cyphers, David Kaufman and Tim Wilson (whose proof reading proved invaluable) for their help and friendship through the last couple of years. Al Sieving, Brad Hines and all the rest at MacGregor get all my gratitude as well for keeping me going. I also thank the rest of the Speech Group for their help, kindness, ready advice, and wonderful working environment.

I also thank Beth, my wife for her love and support without which this this thesis would not have been completed and for whom this thesis was an ordeal as well. Thanks also go to my parents and my whole family for being there, for always being ready to help out and for giving me the chances.

Research support for this thesis has been provided by contracts from DARPA, monitored through the Office of Naval Research, and the Naval Electronics Systems Command.

Contents

Chapter 1

Introduction

As part of her Master's Thesis, Aull constructed a Lexical Stress Determiner for discrete words[1]. As my thesis, I propose to make her system more robust, both from a programmer's point of view and from a performance and reliability perspective.

Aull developed this system in the course of studying the effect of lexical stress information in large vocabulary speech recognition. Her system achieved 87% accuracy in determining the correct number of syllables and the proper stress pattern. Her system was written on a Symbolics Lisp Machine and was designed to interact with the *Spire*[16] speech tool developed at the MIT Speech Group. The system was automated such that you could speak an isolated word to it and it would soon return the stress pattern. Because Aull concluded her work two years ago, extensive updating of her code was needed. The efficiency of the code could also be improved to speed real-time performance. Much of it had to be rewritten in order to run properly on current the Lisp Machine operating system.

Her system consisted of two main components, a syllable detector and a stress determiner. The syllable detector had problems finding boundaries in two cases: 1) when two syllabic nuclei are separated by a sonorant consonant as in "zero" and 2) when there are no intervening consonants, as in "react". Aull's syllable detector was fairly accurate but relatively inflexible. It was not very good at finding syllable boundaries that occurred at Vowel-Voiced-consonant-Vowel transitions. It also had problems with short releases after consonant stops. The stress determiner gave only one answer with no indication of a confidence level. This is a handicap when the system is used as a front end of a large vocabulary lexical access system. If a mistake is made in stress determination, then there is no way to find the correct target group of words. Mistakes can include false insertions of syllables, deletions of syllables and incorrect labeling of stressed syllables. Because the stressed syllables provide "islands of reliability" for acoustic information within the word, it is especially important that the system correctly identify them

The second chapter of this thesis describes lexical stress. It explores what lexical stress is and how it might be important to a speech recognition system. The third chapter describes Aull's system for automatic detection of lexical stress in isolated words, exploring the components of her system developed by others. The fourth chapter explains the modifications that have been made to Aull's system and how they changed system performance. The last chapter contains conclusions and some possible directions for future development.

Chapter 2

The Importance of Lexical Stress

2.1 What is Lexical Stress?

In this paper, as Aull did, I will be dealing exclusively with lexical stress in isolated words. This isolates a stress pattern of the word from higher order effects such as intonation and sentential stress.

Historically, **stress** has been a poorly defined concept. Lexical stress can be described from several points of view. It can be viewed linguistically, phonemically and phonetically. It has been variously described as the force with which a syllable is said or as a feature composed of other features (i.e. fundamental frequency, duration and intensity)[9]. However, it is generally agreed that what we perceive as stress is not a feature of speech (or language) unto itself but is rather a combination of other, more basic, features.

This chapter briefly describes what lexical stress is and then explains some of the motivations for wanting to look at lexical stress and incorporating knowledge about it

into speech recognition systems.

2.1.1 Stress in Language

Stress is a perceived parameter — it is easily detected by a human listener. Most languages have measurable stress effects in their words. In Languages like French, Finnish or Polish, the stressed syllable is fixed on a certain syllable in the word (such as the first syllable or the last one). These languages are said to have fixed stress[8].

Other languages, most notably English, have what is called free stress, meaning that the stressed syllable can fall anywhere in the word. stress can also have higher order knowledge incorporated. In these languages the stressed syllables are not fixed. In these languages, it is words themselves that have stress patterns associated with them. Sometimes the same spelling can have two or more meanings and different stress patterns to go with them (e.g. **"permit"** and **"permit"**). This is especially common when the same word represents a two meanings that are different word types (like in the previous example where permit is first a noun and then a verb).

The difference between stressed and unstressed syllable also changes from language to language[8]. French, for example has very little difference which means that all their syllables are fully articulated. In English, on the other hand, many syllables are not fully articulated, resulting in shortened sonorant regions and schwa's.

In English it is usually true that a word will have a given stress pattern consistently. This is different from other languages where there is either fixed stress in words or there is not enough difference in the stress between syllables to be reliably determined.

2.1.2 Components of Stress

Linguistically, stress is considered a parameter unto itself. The same can not be said from an acoustic point of view. There is no single determiner for stress which means that you can not look at one parameter (energy or some similar measure) which will reliably determine the stress pattern of a word.

The Four Main Correlates of Stress

Through many studies, it has been determined that English stress is primarily determined by four parameters. These parameters are energy, fundamental frequency, duration and phonetic quality[9].

Energy refers to the measure of acoustic intensity of the syllable. Syllables said with more force, exert more pressure on the surrounding air which shows that there is more energy put into the articulation of these syllables. The absolute amount of energy in each syllable is not as important as the energy ratios within the word's syllables. Ratios are more important than absolute values for all these parameters because there is a great deal of variability in speech, not only between different speakers but also different words uttered by the same person[15].

Fundamental frequency, perceived as pitch, is also a main component in the determination of stress. A syllable with higher pitch compared to another syllable, with all else being equal will be heard as the stressed syllable. Many experiments have shown that it is not necessarily the peaks or mean values of the fundamental frequency that correspond to stress perception but rather the shape of the F_0 contour that really matters[9].

Duration is important for stress perception as well. In general, the longer the

duration (relative to other syllables) the more likely that the syllable is going to be perceived as stressed.

Phonetic quality is a measure of how fully articulated the syllable was. Aull measured this parameter when she labeled the qualifying syllables as reduced, that is, they were short and had little energy when compared to other syllables.

How the Correlates Come Together

Using the words in her database, Aull found that no single parameter was a very good indicator of which syllables were stressed. For example, maximal average energy corresponded to the stressed syllable 84% of the time and the peak of the fundamental frequency corresponded to the stressed syllable only 70% of the time. These results were in good agreement with previous data.

Fry[5] found that both duration ratio and energy ratio were important cues for the judgment of stress. He further found that the duration ratio was more reliable than the energy ratio. Morton and Jassem[9] found that changes in fundamental frequency had greater effect on stress perception than did changes in either energy or duration.

2.2 Usefulness of Lexical Stress in Speech Recognition Systems

The obvious question is that of the potential importance of lexical stress in speech recognition systems. We want to know if there is any useful information contained in the stress pattern. For this report, I am limiting my comments to isolated words. When continuous speech is included, higher order stress patterns and rhythm effects

start influencing stress patterns.

2.2.1 Lexical Stress and "Islands of Reliability"

Aull and Zue[2,15], among others, claimed that stressed syllables were reliable places to look for acoustic information. That is, acoustic cues were much more robust in those areas. They further note that spectrogram reading experiments and automatic recognition systems tend to recognize phonemes around stressed syllables more accurately than around unstressed syllables. This result seems to be true in humans as well. Cole and Jakimak[3] found that it took subjects longer to recognize a mispronounced word when the syllable was unstressed compared to when it was stressed.

2.2.2 Lexical Access and Large Databases

After doing studies on a lexicon developed from the Mirriam-Webster Pocket Dictionary, Aull found that lexical stress was very useful in reducing the expected size of word candidates in a recognition system. Studies by Huttenlocher and Zue[6] indicate that determination of broad phonetic classes greatly reduce the number of potential word candidates in an isolated word recognition system. Information about the number of syllables and their stress pattern can augment the phonetic class knowledge to further reduce the word candidates in a recognition system, giving the later (and more detailed) processing of such a system fewer possibilities to investigate.

All the evidence seems to indicate that knowledge of lexical stress would be quite desirable in an isolated word recognition system. The information about stressed syllables points to regions that tend to be more acoustically reliable, improving recognition in those regions. The stress pattern, once determined, also provides an additional con-

straint that can reduce the candidates that a recognition system would have to sift through. Thus determination of the stress pattern is potentially a useful tool in speech recognition systems.

Chapter 3

Aull's Lexical Stress Determiner

3.1 System Overview

Aull's system was designed to determine the stress patterns of isolated words. Her motivation was largely to determine if this would be an effective way to reduce the search for target words in large vocabulary systems. Aull's system was written on a Symbolics Lisp Machine to be used in conjunction with a Floating Point Systems array processor. The system had as an integral component, *Spire*, a speech research tool developed within the MIT Speech Group.

The input to the system was digitized speech with no additional processing, and the output was a time-aligned stress pattern of the word. The time-aligned stress pattern corresponded to the vowel of the syllable and any surrounding sonorant segments. The system labeled the syllables as either "stressed", "unstressed" of "reduced". There could be only one stressed syllable in any word. If two syllables were close in the stress rankings, the system labeled a second choice for the stressed syllable.

The system was broken down into two main sections. The first section was the syllable detector. This section looked for sonorant regions and also looked for intervocalic consonants whose presence indicated a single sonorant region that could contain two or more syllables. The second section was the stress determiner. It performed computations on the different sonorant regions in order to determine their stress ranking.

3.2 The Computing Environment

As mentioned before, this system was developed on Symbolics LM-2's that were equipped with Floating Point System's FPS-100 array processor. The system was built around the *Spire* speech tool as well as including portions of systems developed by others in the Speech Group.

The Computing Environment

The Lisp Machines provided a very flexible and convenient environment in which to work. Both *Spire* and Aull's system made extensive use of a Flavor[1] system which is part of the Lisp Machine operating system. The machine's large virtual memory and networking capabilities allowed the the system to work with a great deal of data. The Lisp Machine also has excellent facilities for system development[16]. The Lisp language itself provided an exceptionally flexible and easy to work in programming environment.

The Lisp Machine has extensive development facilities on which to develop an interactive system. It has very versatile multiple window support and a high resolution bit-mapped display. The speed with which it computes needed parameters also allows

[1]Flavors are structures which are easy to manipulate and facilitate message passing.

convenient interactive research. Using a mouse speeds interaction with the computer and is more "user-friendly" than using the keyboard exclusively.

The system (especially the pitch detector written by Seneff[12]) used the FPS-100 a great deal. The array processor gave a great increase in speed over what would have been possible on the Lisp Machine alone.

The FPS-100 is set up in a master/slave configuration with the Lisp Machine. It sits idle until the Lisp Machine sends it something to do. Chunks of data are assembled in the Lisp Machine and sent out to the array-processor. The array processor then performs a series of steps, or a mini-program (stored there by the Lisp Machine) on the data and finally sends the results back to the Lisp Machine. This continues until the entire waveform (or any array) has been passed through the array processor.

The *Spire* Advantage

Spire is an interactive speech research tool developed at the MIT Speech Group by D. Shipman, D. Scott Cyphers and David Kaufman. It has been evolving for several years and many others have contributed to it.

Spire was developed on Symbolics Lisp Machines, mostly for the reasons stated above. It was intended to be a replacement for other speech tools that existed at the time. Its original implementation by David Shipman was completed in 1982. Following that Cyphers and Kaufman completely rewrote *Spire* in order to make it more flexible, improve the user interface, improve data management and increase its efficiency (both in run-time and in memory usage)[16].

As described by Cyphers[4], *Spire* has a four tier display system. A layout, at the top of the hierarchy, is a screen of data. It is composed of displays, that are like windows.

These displays hold any number of overlays, that are essentially drawing methods. These overlays take on the name of their associated atts. The atts are computations performed on the data and are displayed in the manner specified by the overlay.

Spire works with representations called utterances. An utterance is an event or an instance of someone saying something. That definition, while not very rigorous, is sufficient for my purposes. Attached to the utterance are instances of flavors called attributes. It is the attributes which define the atts.

Spire allows users to easily define new computations and modify old ones. *Spire*'s design allows easy interaction with previously computed data. The display system is the same way; it is very flexible and easily extendable. It is these characteristics that make *Spire* desirable as a speech research tool.

It is a combination of the *Spire* program and the Lisp Machine support that allows systems to be easily built. Since many of the structures and methods needed in a large system are already present in *Spire*, it makes sense and saves work to incorporate it into any system in development.

3.3 Syllable Detection

As I mentioned before, the first section of the system incorporated a syllable detector. Because all syllables must have a vowel at their root, this part of the system attempts to find and separate all the vowel regions in a word. The syllable detector itself has two distinct components. The first is Hong Leung's broad classifier that was developed as part of a system that automatically aligns phonetic transcriptions with continuous speech[14]. The second section, developed by Aull, separated sonorant regions into different syllables if it found any intervocalic consonants.

3.3.1 Leung's Broad Classifier

Leung's broad classifier was the first stage of a system that provided a time-alignment of a phonetic sequence to the speech waveform[7]. Aull used this classifier for her system to obtain a broad segmentation of the speech signal.

The approach that was taken was to first find acoustically robust regions in the waveform. From there, more detailed analyses could be made in appropriate regions that would not necessarily meaningful to make in other regions. This breaks down one large problem into several smaller ones that are more easily approached[14].

The data takes the structure of a binary decision tree. A series of classifiers make decisions about whether or not a time-slice of speech has a certain characteristic. The classifiers are all structurally the same but differ in the parameters that they look at and where they clip their values. The speech is analyzed every 5 msec.

A representative classifier uses **M** parameters, that are decided by previous speech knowledge. The parameters are computed, then processed; they are smoothed, clipped and then normalized. Now, for every 5 msec we have an **M** dimensional feature vector. A decision is made in this **M** dimensional feature space through a K-Means clustering algorithm. In this manner Leung found that he could reliably divide the utterance into six types of regions:

- S (Sonorant) : vowel-like, this would be a syllable core.
- O (Obstruent) : exhibits high frequency "noise".
- VO (Voiced Obstruent) : shares characteristics of both of the above.

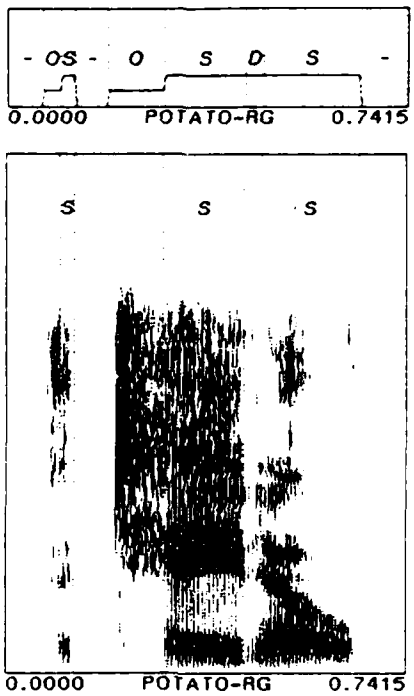


Figure 3.1: An example of the output of the Sonorant Detector including (a) Leung's Region Detector (S means sonorant) and (b) Aull's Syllable Detector (S means syllable).

- Si (Silence) : characterized by absence of energy.
- B (Nasals and voice bars) : similar to sonorants .
- Ul (Unlabeled) : these exhibit energy dips in vowel regions.

The system goes through many classifiers, and segments are re-checked for accuracy and possible low-energy, or weakly represented regions. Decisions can be reversed in later processing to prevent an early mistake from propagating through the entire process.

The segments that Aull was most interested in were naturally the sonorants. This is because they form the root of syllables and hence were the segments that she had to find. Leung's broad classifier was very good at determining boundaries between every type of segments except for different voiced segments. To find harder boundaries (Vowel–Vowel for example), Aull had to develop her own algorithms.

3.3.2 Aull's Intervocalic Detectors

After the initial segmentation by Leung's system, Aull inserted a subsystem that was designed to find intervocalic, voiced regions. This is meant to include both voiced consonants (like the "r" in "miracle") and vowel-vowel transitions (like the "ie" in "anxiety"). These phenomena often exhibit themselves through formant movements or energy dips, but not always.

All of these detectors made extensive use of spectral weighting windows, specifically short-time spectra of the waveform were multiplied by a frequency weighting function designed to bring out spectral characteristics that were expected in certain frequency ranges. Then the results of the multiplication are then accumulated into a Center of Gravity function. The center of gravity function is as follows[1]:

$$Center\ of\ Gravity = \sum_{f=F_1}^{F_2} W(f)\, S(f)$$

where

$W(f)$ = linear weighting window

$S(f)$ = spectrum value at f

F_1, F_2 = frequency range

Although Leung's system performed rather adequately on intervocalic nasals, Aull developed a subsystem just for that purpose. In this part of the system Aull did not choose to use a spectral weighting scheme but rather, she looked for a robust drop in energy in the frequency region of the first three formants (F_1, F_2, and F_3). As a result of Leung's segmenter and Aull's detector the system was quite robust in determining nasal boundaries.

Leung's system was not quite as good at detecting semivowels (/l/, /w/). These are often characterized by a drop in energy similar to the nasals except that only F_2 and F_3 show a significant drop. The drop in energy is more gradual than in the case of nasals.

Even harder to detect were intervocalic semivowels /r/ and /y/. These are characterized by a concentration of energy around 2KHz. There are sometimes dips (at least for /r/) in formant frequencies as well, but far from always. The shape that these semivowels take in the frequency domain are very context dependent and are hence difficult to detect. Leung's system generally misses these completely. Aull used a spectral weighting window that emphasized 2000 Hz and 300 Hz while deemphasizing-emphasizing the frequencies around 1100 Hz. In this way she can label regions as r-like or not r-like[1].

The hardest types of intervocalic activity to detect are the vowel-vowel transitions. For this type of decision, Aull used spectral weighting windows that attempted to emphasize these changes. She took advantage of speech knowledge to determine window that would emphasize transitions between different types of vowels. Even so, these changes are not very robust and are difficult to detect under the best of circumstances.

The syllable detector was designed to identify the sonorant regions of speech for

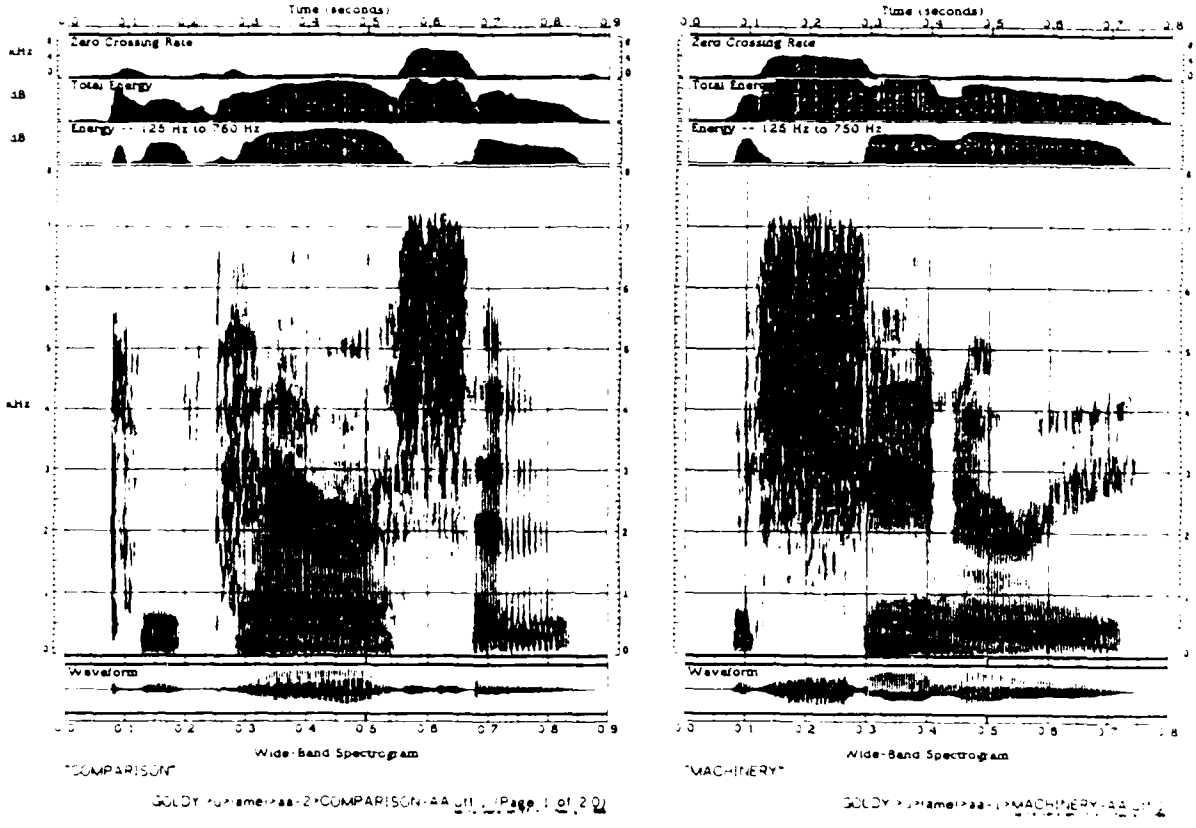


Figure 3.2: Two examples of /r/'s in isolated words (note the differences).

further analysis. Leung's broad classifier first separated the speech into acoustically robust segments. After that, Aull applied a series of processes designed to further separate the sonorant regions by determining if there were any intervening regions between syllable cores.

3.4 Stress Determination

As I mentioned before, it has been found that fundamental frequency (pitch), duration and spectral energy are good correlates of what we perceive as lexical stress. The problem that Aull encountered though, was that any one of these parameters could not determine the stressed syllable correctly more than 87% of the time. As a result she determined that using all of these parameters (as well as one other, spectral change) was more reliable than using any one of them in determining the relative stress of syllables in isolated words.

3.4.1 Acoustic Parameters

One of the parameters that Aull looked at was duration. She used the sonorant region found by the front-end as the basis for her duration measurement. The sonorant boundaries were determined within 5 msec. Any more accuracy would have been unnecessary due to the uncertainties involved with the determination of the boundaries. From her own studies and those by others, she found that the final syllable or sonorant region must have its length adjusted for an effect called prepausal lengthening, i.e. the lengthening of the final syllable in an isolated word.

Aull then looked at the energy over two bands extending from 400 Hz to 5000 Hz

and from 1200 Hz to 3300 Hz. These energies were picked to cover the range of sonorant regions and to deemphasize energy regions associated with consonants.

Fundamental frequency or pitch was the third parameter to be measured. The pitch was determined by using an enhanced waveform, which enhances the fundamental periodicity and then using an Average Magnitude Difference Function of the waveform[12,1]. Aull also mentioned that the peak value of the pitch seemed more significant in determining stress than its average value because of differences between isolated words and continuous speech.

Another parameter that Aull incorporated was spectral change[14]. This parameter was a measure of change of energy in sonorant regions. The energy change was measured across several energy bands according to the following formulas:

$$
\begin{aligned}
S[n] &= \max(D_1[n], D_2[n]) \\
&\text{where} \\
D_1[n] &= \sum_{i=1}^{N} \frac{(E_i((n+1)\,T) - E_i((n-1)\,T))^2}{TE(nT)} \\
D_2[n] &= \sum_{i=1}^{N} \frac{(E_i((n+2)\,T) - E_i((n-2)\,T))^2}{TE(nT)} \\
TE(nT) &= \text{total energy of entire spectrum} \\
E_i(nT) &= \text{energy value in ith energy bank} \\
T &= 5 \text{ msec}
\end{aligned}
$$

This parameter was used because it was found that stressed syllables were more acoustically stable than unstressed ones. This parameter was only extracted in the

central parts of the sonorant regions so that the surrounding regions could not influence the spectral change measurement.

3.4.2 Stress Determination Algorithm

Aull had to combine these parameters into one meaningful measurement of stress. She initially tried K-means Clustering techniques but found that they did not perform adequately. The main problem with any system that looks across a group of words is that there is too much variability across isolated words. She then dropped this and other methods that required accumulating statistics across many instances of isolated speech and instead adopted a method that used only the particular word that the system was currently working on.

She associated a five-dimensional feature vector with each sonorant region. Then, for each parameter, the system determined the maximum value across all the sonorant regions and collected them into a maximum feature vector. This maximum feature vector was the basis to which the sonorant regions in the word were compared. This reduced interword variability.

A Euclidean distance from the maximal feature vector to each sonorant feature vector was determined. The region with the shortest distance was considered to be the stressed syllable. The other syllables in the word were all labeled unstressed. Further processing determined which sonorant regions were reduced by looking at their energy and duration.

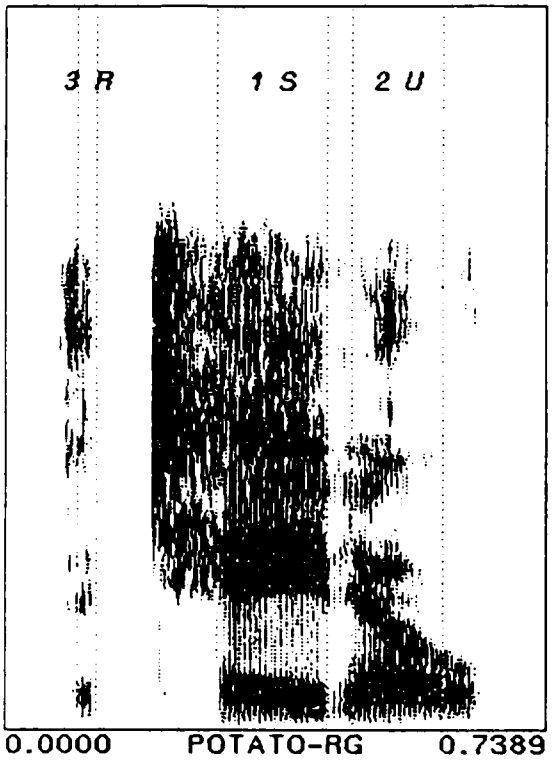


Figure 3.3: An Example of Stress Determiner Output. The numbers correspond to ranking and the letters mean Stressed, Unstressed or Reduced.

3.5 System Performance

Aull tested her system on a 1600 word database. Her system correctly determined the stress pattern 87% of the time. 3% of this error was due to confusion between unstressed and stressed syllables within the word. The other 10% corresponded to either missing a sonorant region, failing to insert a boundary in the case of intervocalic phenomena or false insertion of a region or boundary.

Aull determined that the acoustic correlates of lexical stress, as determined by Fry[5] and others, were quite adequate for determining the stress in a word. She did find, much as she expected, that her system performance degraded as acoustic cues became more subtle.

3.6 Summary of Aull's System

Aull's system consisted of two main subsystems, a syllable detector and a stress determiner. The syllable detector was made up of Leung's acoustic front end and Aull's intervocalic detectors. The stress determiner extracted a five-dimensional feature vector from the sonorant regions. These parameters have been experimentally determined to influence perception of stress. The feature vectors were then compared to a maximal vector for stress determination.

Chapter 4

Modifications of Aull's System

4.1 System Flaws

Aull's system was very good but, like all systems of its type, was not perfect. Aull measured the system at 87% accuracy. That figure refers to correct determination of the syllables and the stress pattern. She found that 3% of the time, the stress pattern was not determined correctly by the system. This means that 10% of the time, there was a problem in finding the syllables correctly. These errors correspond both to false insertions and false deletions.

Thus the largest problem that the system had was in the area of syllable detection. This part was difficult because it relies on acoustic cues, some of which can be quite ambiguous. The stress determiner, while not perfect, is more robust than the syllable locator because the parameters used to determine stress have been heavily studied and are fairly well understood. While this section also relies on acoustic parameters, it is constrained to the boundaries determined by the syllable detector.

One of the biggest problems was that Aull's system was written almost two years ago. Much of what she had done was unexplained. The computers that the group currently works with are different from the ones that Aull worked on (though they were still Symbolics machines and retained a great deal of compatibility). Also, both the Lisp Machine operating system and *Spire* had undergone several major changes. These conditions added up to the fact that much of the existing code had to be changed in order to get the system running as before.

4.1.1 Problems with Syllable Detection

Aull's system, as I stated before, was evaluated at about a 10% error rate for the syllable detection section of the system. The system performed quite well in identifying syllables that are separated by obstruents (as in "duplicate") . These boundaries were correctly determined by the acoustic front end and required little additional processing.

The syllable finder's performance decreased as the consonantal regions between vowel regions became less obstruent-like. This lowering of performance is due to the fact that some intervocalic voiced consonants appear more vowel-like than others. As mentioned before, /r/'s,/l/'s are always hard to find, because sometimes they take on vowel-like acoustic properties.

The system also had trouble with vowels whose amplitudes are low. This phenomenon occurs in reduced vowels. Some people reduce them more than others and sometimes the reduction results in a deletion of the region. Even when there was a very short, low energy sonorant, a human listener will still detect a syllable there. The solution is to detect these regions and then eliminate any false alarms resulting from from making the system more sensitive.

Finally, the system has a great deal of difficulty with syllables that are not separated by consonants due to the fact that the acoustic cues for vowel-vowel transitions may be subtle and are not well understood.

4.1.2 Problems with Stress Determination

While, this component proved to be more reliable than the syllable locator, it still had several problems. The largest problem was that it was not flexible enough for lexical lookup into a large lexicon. The system provided a stress pattern and no additional information. This means that the system will either be right or wrong, there is no margin of error. There is no second choice or quality of decision information. An improvement in this part of the system would allow more flexibility and would go a long way to remedying the problem of misidentifying a word's stress pattern.

4.2 System Code Changes

Almost two years elapsed between when Aull finished her research and when I started to look into her system. Unfortunately the system and machine that that her stress determiner ran on did not remain static through that time. The Speech Group updated its machines to the newer Symbolics 3600 Series Lisp Machines, Symbolics also introduced numerous changes in its operating system, and, most significantly, *Spire* was extensively rewritten by D. S. Cyphers and David Kaufman. All these changes contributed to the work that had to be done in order to return the system to its former status and hopefully beyond.

4.2.1 Updating System Code

The first thing to be done was to rewrite the code so that it would run again. Getting a system that would run and one that would run correctly turned out to be two different things. To get the system running mostly entailed recompiling the system and making some simple changes that included changing message names and other system updates.

Much of the system's *Spire* interface had to be rewritten in order to run properly. Since Aull had finished her work, *Spire* had changed a great deal. Both *Spire* displays and the representation of time-aligned data had changed incompatibly. In both cases (operating system and *Spire*) there were also subtle changes that affected system performance. These had to be corrected individually as they were found.

4.2.2 Improving System Efficiency

Improving system efficiency and run-time performance was a different issue from updating the code. After the code had been updated, it was found that there were many places that would benefit from being rewritten or modified. Some of the modifications were for the sake of computation efficiency and others were done in order to make the code more compact and smoothly flowing. The biggest change that were made had to do with the way in which segments and their boundaries were accessed.

The major running-time improvement was contributed by Seneff who wrote a version of the Gold-Rabiner pitch detection algorithm[11]. This algorithm was much faster than the algorithm then being used. It seems that the system changes introduced by this author have also improved the run-time performance of the system but it is difficult to substantiate. The speed of the system was further improved by numerous hardware

modifications to the Symbolics machines. The system now runs at least three times faster than it did before, or about 95 times real time.

4.2.3 Improvements in System Flexibility

The system as Aull left it was rather rigid in that parameters that went into computations could not be modified or accessed easily. Modification of computational parameters is a task that *Spire* makes easy. What was done was to make these parameters changeable from *Spire* so that it was not necessary to constantly recompile the system code when changing numbers or parameters. This change facilitated the development stage, when thresholds were specified iteratively in order to minimize both false insertions and deletions of segments.

4.3 Changes in Syllable Detection

The syllable detection section was broken into two different parts for the purposes of modification. They followed the natural division of this section, that is Leung's front end and Aull's detectors for intervocalic events.

4.3.1 Improving Sonorant Detection

The first goal was to improve the sonorant detection in the acoustic front end. This is an important step because if a sonorant region was not detected there, it would be unavailable for of all subsequent processing. However, erroneously inserted segments arising from making the system more sensitive to sonorant regions could be eradicated in later system components.

The front end was missing sonorant regions, but was also falsely breaking up valid regions. That is, it would first label a region sonorant and then insert another label in the middle of it. This had the result creating two invalid sonorant regions from one good one.

The solution to resolving the missing sonorant problem was found in the K-Means clustering algorithm used in the system. The initial step in that algorithm was to establish clipping values for each of the parameters investigated. The clipping values were the extremes in the data-space that a given type of region was expected. These clipping values were reevaluated iteratively so that bad regions were minimized, while low amplitude sonorants were maximized. The effect was that many low energy sonorants were found, while few false insertions resulted.

In order to decrease the number of false insertions into the middle of sonorant regions, some of the processing done in the front end had to be eliminated. The system would initially find and label sonorant, obstruent and silent regions. It would then segment the regions further by looking for different acoustic cues within these regions. It is in this later processing that the errors (the false insertions into the sonorant regions) usually occurred. The key to solving this problem was to determine at what point in the processing the most errors were inserted while not missing too many valid regions. It was determined that some processing after the initial labeling was necessary in order to keep the number of false insertions down to a minimum.

After changing the front end, more problems had to be dealt with. First, many of the new sonorant segments were discarded by a module that tried to decide what was really a sonorant and what wasn't. This part of the system looked at the duration, energy and spectral change of the sonorant region, and if the region was too short

and/or had too little energy, it was deleted from the syllable list.

The thresholds at which the system would cut off sonorant candidates was changed iteratively. Once the levels were changed, it had to be ensured that the falsely labeled regions were kept to a minimum while the truly sonorant regions detected were maximized. In the end, all the average energy thresholds needed to qualify a segment as sonorant were lowered. At the same time it was determined that the average length of the falsely detected segments was less than even the most reduced real sonorant regions. Because of this, I lowered the durational threshold as well. This allowed the low energy sonorants to be detected while still keeping the false indications to a minimum.

Spectral change is used as a parameter in this computation because Aull felt that if a region exhibited a great deal of spectral change then it was less likely to be a sonorant than a more spectrally static segment. This conclusion is not exactly obvious for segments of such short duration, but the inclusion (or modification) of this parameter has not caused any system deterioration. Because the spectral stability gives another clue to the segment's identity, impostor sonorants of greater duration can be more reliably removed from consideration.

4.3.2 Improving Detection of Intervocalic Consonants

This is the section that proved to be the most disappointing in terms of improving system performance. While it seemed to make moderate performance gains in /l/ detection through changing some thresholds it had greater difficulty with /r/'s.

The problem was that if the system were made more sensitive to the spectral movement that often occurs with intervocalic /r/'s (as in "interrupt"), it would then get more false boundary insertions at /r/'s that were not intervocalic (as in "cohort").

This trade-off between insertions and deletions was a problem that had always plagued the system. Maybe a different parameter would have been better but time did not permit an investigation of this possibility. One parameter, a spectral first difference, was investigated but preliminary results indicated that it was not useful for intervocalic /r/ detection.

Another difficult problem was that of vowel-vowel transitions. Detecting these transitions reliably is difficult because of the many different, often subtle changes they produce in the spectrum of the word. Sometimes the changes can be very obvious while at other times, they can manifest themselves through slow formant changes. The difficulty in finding and interpreting them is compounded by their variability from speaker to speaker.

The previous two problems received much attention, mostly in the form of changing parameters and thresholds, both in the acoustic front end and in Aull's intervocalic detectors. Unfortunately both met with little success.

4.4 Changes in Stress Detection

The biggest problem with the stress determination mechanism was that it was not very flexible. In addition it sometimes failed to correctly find the stressed syllable all the time. A large part of this second problem can be attributed to the variability with which sonorant regions surrounding the vowel are included in the segment. All other things being equal, the region that is longer will be considered stressed by the system. This could be a problem when two regions are similar in the amount of stress that can be attributed to them and one region is significantly longer than the other. Different weighting functions for the parameters in the distance metric were tried but

this provided no measurable change in the stress determination.

To improve the flexibility of the system, new methods to provide data for further processing of the stress information were considered. Time did not permit proper investigation of the usefulness of the results of these methods, but it is felt that they could contribute to overall system performance. Also, it was felt that this section was not as critical as others because this aspect of the system performed relatively well.

a way to obtain the actual measurements of the system (rather than just "stressed" or "unstressed") was provided . This allows one to look at results of the Euclidean distance measurement across the **M** dimensional feature vector. Another addition that was made (and kept in the system because it both provided additional information and was easily interpreted) was the inclusion of the ranking of the syllables rather than just labeling them "stressed", "unstressed" or "reduced". This allows the user to see what the output of the system is more clearly.

4.5 System Evaluation

The system was tested on 228 isolated words spoken by six speakers (3 male and 3 female). These words were taken from databases used by Aull. Her system returned errors on all these words at some point. Some of the words were evaluated correctly by her final system but were included to determine if changes to the system degraded performance on data that was already valid.

In Table 4.1, the V-V, Cons, and Son columns all correspond to a miss in either the vowel-vowel(like in "anxiety"), consonant (such as /r/ or /l/) or sonorant (as in the last syllable of "action") contexts. The Insert column refers to false insertions of sonorant regions and the Bad Stress refers to incorrect stress assignment. The numbers

Table 4.1: Evaluation Results

System	V-V	Cons.	Son.	Insert	Bad Stress
Original	24	46	37	63	7
Modified	24	47	22	56	5

are total number of that type of error resulting from evaluating the data base. This was done because some words resulted in more than one error while others resulted in none.

The number of missed vowel-vowel transitions did not change at all. This was expected because nothing was done to the system that would directly affect performance here. The important thing is that system performance did not degrade. The same can be said for the missed sonorant regions. Although, an effort was made to improve performance in this area, it was unsuccessful.

The number of missed sonorant regions dropped significantly. This was due mostly to the changes in the initial processing of the acoustic front end. The remaining undetected sonorant regions were very short and had low energy but still could be perceived as syllables to human listeners.

The number of incorrect insertions also dropped. There were two effects going on in this case. The sonorant detector defined more regions as sonorant than it did before because of its increased sensitivity. That increased the number of false insertions. On the other hand, fewer valid sonorant regions were being broken up, driving the number of false insertions down. This was the dominating effect, bringing the total number of

insertions down.

The number of words with incorrect stress assignment also dropped slightly. This was due to the fact that sonorants had different boundaries than before, hence the measurements for those regions were different. This was the only effect taking place since there were no computing changes made to the system. It was found, however, that in every case of bad stress assignment, the stressed syllable was always ranked second, showing that the system was close.

In the course of investigating these results, it was found that most problems could be corrected interactively. This indicates that system performance might be able to benefit from some sort of time varying evaluations on a frame by frame basis.

4.6 Summary of System Improvements

The changes made to Aull's system led to several improvements. These improvements are:

- Run Time Performance - The speed of the system through improvements in code efficiency and hardware changes decreased running time three fold to about 95 times real time.

- System Flexibility - Through code changes, the system was made easier to use and change interactively.

- Syllable Detection - The system detected syllables more accurately through changes that improved identification of sonorant regions. In addition, the insertions of spurious regions into the middle of valid vowel regions is reduced. In

these two areas, the number of errors was reduced from 37 to 22 and from 63 to 56 (40% and 11%) respectively.

System performance did not degrade in any way as a result of these changes.

Chapter 5

Conclusions

5.1 Summary

In this thesis, lexical stress was described and its potential utility in automatic speech recognition was outlined. The stress is a perceived quality measure of a syllable. Acoustically stress can be determined primarily through four parameters: spectral energy, duration, fundamental frequency and spectral quality. The lexical stress pattern of a word is useful to determine in an automatic recognition system because it reduces the search space of possible candidates.

Next, the system that Aull developed for her Master's Thesis was investigated. The system was made up of two main parts: A syllable detector and a stress determiner. The syllable detector was composed of an acoustic front end and a series of intervocalic consonant detectors. The stress determiner took an **M**-dimensional feature vector of each sonorant region and compared it to a maximum feature vector and from that the syllables were ranked.

Then the problems that existed in Aull's system were explained. These included poor performance on some intervocalic consonants, on Vowel-Vowel transitions and on low amplitude sonorants.

Finally this author's changes to the system were described. These changes to Aull's system did improve its performance. The system ran approximately three times faster, had improved sonorant detection, had fewer false insertions and was more flexible. The number of missed vowel regions decreased by 40% and the number of false insertions into sonorants decreased by 11%. Other regions were not improved, as indicated by the evaluation data, but in no case did system performance deteriorate.

5.2 Suggestions for Future Research

There are many ways to further improve on the work done so far on this system. Many of the parameter values used in the system (that have been changed by this author) can still be improved on using statistical tools and knowledge of speech signals and production. A different method for detection of intervocalic effects, also utilizing more speech knowledge, could also be incorporated. More improvements in code efficiency could be made, to be sure. An algorithm for assigning probabilistic values to stress rankings would be quite useful for making the system suitable for incorporation to an isolated word recognizer.

Bibliography

[1] Aull, Ann Marie. *Lexical Stress and its Application in Large Vocabulary Speech Recognition.* S.M. Thesis, MIT, August 1984.

[2] Aull, Ann Marie and Victor W. Zue. *Lexical Stress and Its Application in Large Vocabulary Speech Recognition.* Presented at ICASSP 85, Tampa, FL, March 26–29, 1985.

[3] Cole, Ronald A. and Jola Jakimak. *How are Syllables Used to Recognize Words?* The Journal of the Acoustical Society of America, Volume 67, Number 3, pp 965 - 970, March 1980.

[4] Cyphers, D. Scott. *Spire Reference Manual.* MIT Speech Group, Massachusetts Institute of Technology, 1986.

[5] Fry, D. B. *Duration and Intensity as Physical Correlates of Linguistic Stress.* The Journal of the Acoustical Society of America, Volume 27, Number 4, pp. 765 - 768, July 1955.

[6] Huttenlocher, Daniel and Victor Zue. *A Model of Lexical Access from Partial Phonetic Information.* Presented at ICASSP 84, San Diego, CA, March 19–21,

1884.

[7] Leung, Hong Chung. *A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech*, S.M. Thesis, MIT, January 1985.

[8] Malmberg, Bertil. *Phonetics*. Dover Publications, Inc., New York, 1963.

[9] Morton, John, and Wiktor Jassem. *Acoustic Correlates of Stress*. Language and Phonetics, Volume 8, pp. 159 - 181, 1965.

[10] Oppenheim, Alan V. and Ronald W. Schafer. *Digital Signal Processing*. Prentice-Hill, Inc., Englewood Cliffs, New Jersey, 1975.

[11] Rabiner, Lawrence R., and Bernard Gold. *Theory and Applications of Digital Signal Processing*., Prentice-Hall, Inc., Englewodd Cliffs, NJ, 1975.

[12] Seneff, Stephanie. *Pitch and Spectral Analysis of Speech Based on an Auditory Model*. MIT Ph.D. Thesis, January, 1985.

[13] Shipman, David W. and Victow W. Zue. *Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems*. Presented at ICASSP 82, May 1982.

[14] Zue, Victor W. and Hong C. Leung. *A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech*. Presented at ICASSP 84, San Diego, CA, March 19–21, 1984.

[15] Zue, Victor W. *The Use of Speech Knowledge in Automatic Speech Recognition*. Invited paper for IEEE Special Proceedings on Man — Machine Speech Communications, November 1985.

[16] Zue, Victor W., et al. *Spire: The Development of the MIT Lisp-Machine Based Speech Research Workstation.* Presented at ICASSP 86, Tokyo, Japan, April 8–11, 1986.

From the Proceedings of ICASSP 86, the IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, held in Tokyo, Japan, April 8-11, 1986.

LEXICAL ACCESS AND VERIFICATION IN A BROAD PHONETIC APPROACH TO CONTINUOUS DIGIT RECOGNITION*

FRANCINE R. CHEN**

Department of Electrical Engineering and Computer Science
and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper describes an implementation of a robust method of lexical access and a detailed phonetic verification component for recognizing continuous digits using a broad phonetic approach. The lexical access component uses a scoring method which takes into account soft labeling errors due to input signal variability. Verification is based on the use of a small set of detailed acoustic features which characterize phone hypotheses. Evaluation of the lexical access method on a database of 74 new random length digit strings, each spoken by 5 new speakers, shows the method to be tolerant to front-end errors and variations in pronunciation. Evaluation of the verification component indicates that use of a few detailed phonetic features is adequate for verification of phones in the digit vocabulary.

INTRODUCTION

In ICASSP-82, Shipman and Zue [1] showed that a broad phonetic representation imposes strong sequential constraints on words in the English language. They then proposed an isolated word recognition model which uses the constraints provided by a broad phonetic representation. In their model, the speech signal is segmented and classified into several broad categories which can be determined reliably. Next, indexing into the lexicon, only words which match the sequence of broad phonetic labels remain as contending word candidates. Finally, the contending words are examined using detailed phonetic analysis to identify the input utterance.

Chen and Zue [2] extended Shipman and Zue's isolated word recognition model to continuous speech and showed that strong lexical constraints at the broad phonetic level can be exploited in a continuous digit recognition task. To illustrate that the approach is viable, a broad phonetic classifier and lexical access component were implemented. Testing on 1718 digits by 5 speakers, the correct digit was not one of the lexical candidates only 1% of the time. While the results were encouraging, this initial implementation suffered in one important respect: The implementation did not provide flexibility in accommodating similar but new acoustic realizations of a word. Instead, new pronunciations were accomodated by explicitly adding them to the lexicon. In other words, a digit was considered a candidate only if the input string was a pronunciation supplied by the lexicon.

In the current study, two aspects of the broad phonetic recognition model were focused on. First, Chen and Zue's work was extended in an effort to develop a more robust method of lexical access which could tolerate reasonable "errors" by the broad phonetic classifier. Second, a preliminary examination of verification of word hypotheses based on detailed phonetic features was performed.

LEXICAL ACCESS

Researchers (e.g. [3] and [4]) have developed systems which perform lexical access and recognition directly from a phonemic sequence. In contrast, this study is based on the belief that a more robust recognition method is to perform lexical access by scoring how well the *broad* phonetic representation of an unknown utterance matches the phonetic representation of a word in the lexicon. Since less detailed distinctions are needed to produce a broad phonetic representation than a detailed phonetic representation, one should be able to compute a broad phonetic representation with less error.

In the broad phonetic recognition model (Figure 1), the broad phonetic classifier produces a broad class segmentation string of the incoming signal. The segmentation string may be composed of six possible labels: weak fricative, strong fricative, short voiced obstruent, vowel, sonorant, and silence. The lexical component matches the phonetic representation of each word in the lexicon against the broad class segmentation produced by the system, yielding a lattice of word candidates.

Although a broad phonetic representation is more robust than a detailed representation, unanticipated acoustic realizations do occur, resulting in classification errors at the broad phonetic level. For example, the closure in a stop gap may be incomplete, resulting in a "noisy" stop gap which is labeled as a "weak fricative". A lexical access component was implemented which attempts to handle these labeling errors using two types of knowledge:

* This research was supported by the System Development Foundation, a Vinton-Hayes Fellowship, and DARPA under contract N00014-82-K-0727 as monitored through the Office of Naval Research.

** F.R. Chen is now with Hewlett-Packard Laboratories, Palo Alto, CA.

 CH2243-4/86/0000/1089 $1.00 © 1986 IEEE

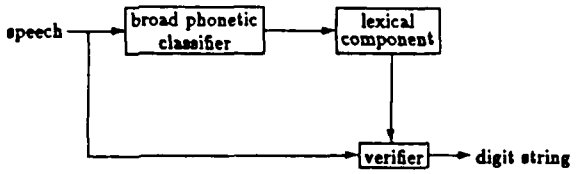


Figure 1: Broad phonetic recognition model

1) how often a phoneme is mislabeled as another class—for example, how often a /k/ closure is labeled as a "weak-fricative" instead of "silence" and 2) how often an insertion or deletion occurs in mapping a word's broad class representation to a phonetic representation—for example, the frequency with which /s/ and /n/ in the sequence /sn/ (as in "six nine") are both labeled as "strong-fricative," due to the fact that the initial nasal in that context may be deleted or extremely short. By using these types of knowledge about the characteristics of the segmentation strings produced by the front-end, the lexical component allows for acoustic variations in a phone. Furthermore, many alternate broad phonetic representations of a word needed with the explicit matching method become unnecessary.

The lexical component assigns a score reflecting how well the phonetic representation of a word matches a portion of the segmentation string, using knowledge about the characteristics of the broad phonetic classifier's output. For example, the broad phonetic classifier may label /θ/ as "weak fricative" 60% of the time and "strong fricative" 40% of the time. Knowing this, the lexical component does not penalize the score much when matching /θ/ to "strong fricative." In contrast, if /θ/ is never classified as "vowel" during training, then the match of /θ/ to "vowel" would be assigned a poor score. Insertions and deletions are handled by using transition probabilities. If the broad phonetic classifier consistently misses prevocalic nasals, as in the word "nine", then the system will know that very often the /n/, as well as the /aʸ/, is labeled as "vowel". This is reflected by a high transition probability of matching /n/ to "vowel" and then matching /aʸ/ to "vowel".

A forward dynamic programming algorithm finds the best match between the broad phonetic and phonetic strings. Simple slope constraints require the path to be non-decreasing in each direction. In contrast to the constraints used in dynamic time warping of the speech signal, many phonetic labels may map into a single broad phonetic segment. For example, the /I/, /r/ and /oʷ/ in "zero" may map into the label "vowel" if the broad phonetic classifier has no knowledge for differentiating among these sounds.

The allowed paths from a sample node are illustrated in Figure 2. Each node represents the match between a broad phonetic label and a phone. The sequence of broad phonetic labels (reference) aligns with the nodes from left to right; and the sequence of phones (test) aligns with the nodes from top to bottom. Three paths, or transitions, exit from a typical node, here labeled "A". When no insertion or deletion occurs, the next broad class segment is matched to the next phone label; this is represented by Path AC. If an extra label is created by the front-end, an insertion occurred; this is represented by Path AB. And if the broad phonetic classifier labels two sequential phones as the same broad phonetic class, a deletion occurred; this is represented by Path AD.

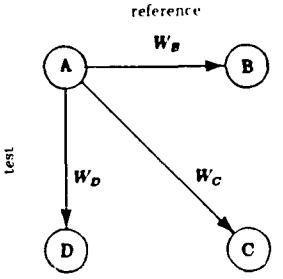


Figure 2: Paths used in the dynamic programming algorithm

The total accumulated score to node C, d_C, is:

$$d_C = d_A + \log[\Pr(p_C, l_C) * W_C]$$

where d_A is the total accumulated score to node A. $\Pr(p_C, l_C)$ is the probability of labeling the phone at node C, p_C, as the broad class label l_C. W_C is the probability of making a transition from node A to C, given that node A is the current state and nodes B,C, and D are states which may be entered from node A. W_C is computed as:

$$W_C = \frac{\Pr(p_A l_A \rightarrow p_C l_C)}{\Pr(p_A l_A \rightarrow p_B l_B) + \Pr(p_A l_A \rightarrow p_C l_C) + \Pr(p_A l_A \rightarrow p_D l_D)}$$

W_B and W_D are computed similarly and represent, respectively, the probability of inserting and deleting a segment.

The "best" alignment between the phonetic string /zIroʷ/ and the broad phonetic representation "strong-fricative vowel" is shown on the left of Figure 3; the associated match and transition probabilities are shown on the right. A phonetic string is assigned the score of the best path, normalized by the number of transitions in the path.

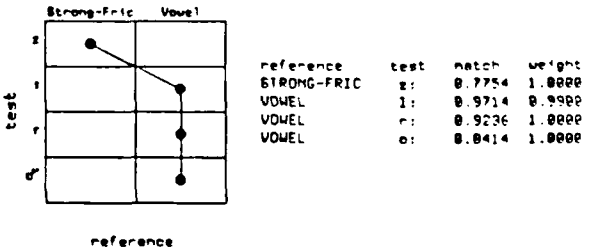


Figure 3: Alignment of /zIroʷ/ with "strong-fricative vowel"

This method of lexical access was evaluated on a database of digit strings ranging in length from one to seven. The database was subdivided into training and new speakers and into training and new sentences, resulting in four mutually exclusive subsets as shown in Table 1.

Table 1: Corpus Subsets

Total # of Utterances	Speakers	Total # of Digits
262 training	3 male, 3 female: training	1365
152 training	3 male, 2 female: new	599
370 new	2 male, 3 female: training	1440
370 new	3 male, 2 female: new	1440

Each broad class segment produced by the broad phonetic classifier was used as the beginning segment of each word hypothesis and the scores of all possible matches were computed. The distributions of scores for *correct* words and for *all* word hypotheses, evaluated on new utterances by new speakers, are shown in Figure 4 as dashed and solid lines, respectively. Note that the log probability scores of the correct words are much closer to 0, or a probability of 1, than the bulk of the scores of all possible words. The distributions indicate that a word score threshold can be set such that all words with a score below the threshold can be ruled out as viable candidates.

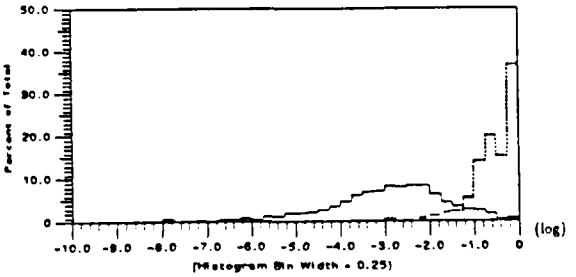


Figure 4: Histograms of correct and incorrect word scores

Figure 5 illustrates the relationship between the amount of pruning achieved compared to the percentage of correct words pruned when evaluated on new utterances by new speakers. Note that one can reduce the number of hypothesized words by 50% without pruning any of the correct words. The curves for training and new speakers were found to be similar [5], indicating that the method is potentially robust to speaker variabilities.

VERIFICATION

In the broad phonetic recognition model, the input to the verifier is a lattice of word candidates produced by the lexical component, the most unlikely candidates having been removed. The verifier selects the best word or string of words from among the competing word candidates using a set of detailed acoustic features.

Each word hypothesis is represented as a sequence of phones and each phone is characterized by a set of detailed acoustic features. This choice of representation was motivated by linguistic reasons and by the desire for extendability to other recognition tasks.

Observations of phone characteristics in spectrograms were used to select a small set of nine acoustic features. These features were designed to capture salient acoustic

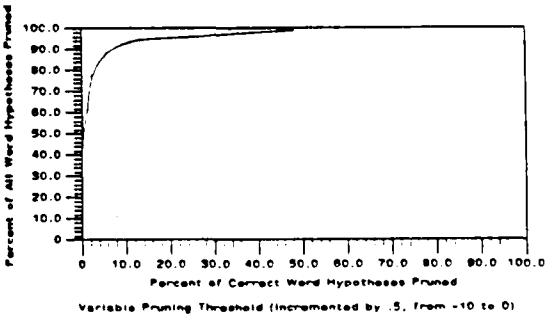


Figure 5: Pruning of all word hypotheses versus correct word hypotheses

characteristics of speech sounds and detailed differences between similar phones in the digits. The features are:

- Position of the first three formants and movement of the first two formants: In an effort to achieve a robust characterization of formant motion and position, a gross characterization based on spectral weights was used, rather than a formant tracker which can exhibit inconsistent behavior in nasalized regions. The spectral weights emphasize energy in specific regions of the spectrum.
- Nasal possibility: To detect the presence of the low frequency resonance characteristic of nasal murmurs, this feature compares the energy in a passband of 100–350 Hz to energy in a passband of 350–850 Hz.
- Onset rate: This feature is the maximum change in energy from 1000–7000 Hz within 20 msec of the beginning of a phone. To capture rapid transitions, the energy is computed every msec from the short time Fourier transform using 2 msec Hamming window.
- Spectral offset location: This feature represents the location of the first spectral dip higher in frequency than the first major concentration of energy in a smoothed spectrum.
- High frequency energy change: This feature is the slope of the best linear fit to the energy in the 4500–7800 Hz band over the duration of a phone. This feature is intended to help differentiate between fricatives (which have relatively stable energy) and unvoiced plosive releases (which generally have a strong onset followed by aspiration which weakens).

Hypotheses scoring can be viewed as a discrimination or identification problem. A binary discrimination allows small differences between similar candidates to be weighed. In contrast, identification indicates how well the measured feature values match the expected values for a phone, independent of the values for the other phones. Because lexical access based on a broad phonetic representation results in similar sounding word candidates, the sounds to be scored should be similar; hence discrimination seems the better approach. Preliminary results bear out this expectation, and a metric based on discrimination between competing phones was used in scoring [5].

To identify errors due to the verification algorithm, the inputs were idealized by mapping the phonetic transcription of each utterance into a broad phonetic

transcription and then performing lexical access on these "ideal" broad phonetic transcriptions. The verifier was evaluated only on the subset of the lexical access database which was phonetically transcribed.

Table 2 shows the word error rates under various test conditions. Each insertion, deletion, or substitution was counted as an error. The error rates illustrate the power of using a few carefully selected acoustic features combined with statistical measures to score each contending phone. On new utterances, the error rate for training speakers is only slightly better than for new speakers, indicating that an acoustic-phonetic approach is potentially speaker-independent.

Table 2: Word Error Rates

Utterances	Speakers	# of Speakers	# of Digits	Word Error Rate
training	training	6	1365	1.5%
training	new	3	364	4.9%
new	training	4	1126	5.0%
new	new	4	893	5.3%

Detailed analysis of the errors in all corpora revealed that many of the errors were due to differences in male/female speech. The most striking and consistent error was the confusion of "four" and "five". All 16 cases in which "five" was incorrectly recognized as "four" occurred in speech by males. Eighteen of the 19 cases in which "four" was confused as "five" occurred in speech spoken by females.

To obtain an indication of the robustness of the verification scores, the score of the correct word relative to the score of competing candidates was examined. When the top candidate was correct, its score was compared to the second best candidate's score. When the top candidate was incorrect, its score was compared to the correct word's score. Figure 6 shows the results of evaluation on new utterances by new speakers. Note that the difference in word scores is generally small when an incorrect word is the best scoring word (dashed line), and that the difference has a large range when the correct word is the best scoring word (solid line). In a recognition system, this information could be used to identify words which do not score much better than their competitors. Finer analyses could be performed on these words or the word could be rejected.

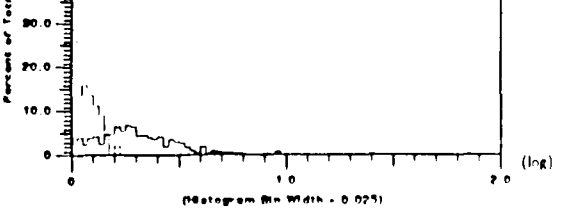


Figure 6: Difference in word scores for correct and incorrect classification

The rank of the score of each phone in the correct word is shown in Table 3. Note that for new utterances by both the training speakers and the new speakers, the correct phone is in the top position at least 86% of the time and within the top two candidates at least 98% of the time. This similarity in rank again indicates the potential speaker-independence of using acoustic features in verification.

Table 3: Phone Rank in Correct Words

Utterances	Speakers	Position			
		0	1	2	3
training	training	93	99	1.00	1.00
training	new	90	99	99	1.00
new	training	86	98	1.00	1.00
new	new	86	98	99	1.00

SUMMARY

Two components of a broad phonetic based continuous digit recognition system have been examined. A method for lexical access was implemented and shown to allow a recognition system to tolerate reasonable front-end variations in labeling. The use of a small set of fine phonetic features for word verification was investigated and found to be adequate. Additionally, evaluation showed these components to be potentially robust to speaker variations. These results are encouraging and indicate that a broad phonetic approach is viable, but evaluation should now be performed on a larger database.

ACKNOWLEDGEMENT

The author thanks Professor Victor Zue, who strongly influenced this work by providing many ideas and much support.

REFERENCES

[1] D.S. Shipman and V.W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proc. ICASSP-82*, pp. 546-549, 1982.

[2] F.R. Chen and V.W. Zue, "Application of Allophonic and Lexical Constraints in Continuous Digit Recognition," *Proc. ICASSP-84*, pp. 35.3.1-35.3.4, 1984.

[3] S. Makino and K. Kido, "A Speaker Independent Word Recognition System Based on Phoneme Recognition for a Large Size (212 Words) Vocabulary," *Proc. ICASSP-84*, pp. 17.8.1-17.8.4, 1984.

[4] K. Shirai and T. Kobayashi, "Phrase Speech Recognition of Large Vocabulary Using Feature in Articulatory Domain," *Proc. ICASSP 84*, pp. 26.9.1-26.9.4, 1984.

[5] F.R. Chen, "Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary," Ph.D. dissertation, Mass. Inst. Tech., Cambridge, MA, 1985.

From the Proceedings of ICASSP 86, the IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, held in Tokyo, Japan, April 8-11, 1986.

A BROAD PHONETIC CLASSIFIER

Daniel P. Huttenlocher

Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

It has been shown that broad phonetic sequences partition a large lexicon into small equivalence classes of words sharing the same sequence. While these results illustrate the power of broad phonetic constraints for differentiating words from one another, they do not suggest how to exploit sequential constraints in recognition. This paper presents a method for decoupling sequential phonetic constraints from a lexicon, by representing allowable broad phonetic sequences in terms of n-th order Markov models. A simple frame-based broad phonetic classifier is used to evaluate the effectiveness of these models in recognition. Tests on 300 sentences from 30 male speakers demonstrate that the addition of sequential constraints improves the classifier's performance.

INTRODUCTION

We have been investigating the use of broad phonetic sequences for hypothesizing words in speech recognition [1]. Shipman and Zue [2] demonstrated that broad phonetic sequences are powerful for discriminating among the words in a large lexicon. They showed that a large lexicon can be partitioned into small equivalence classes by representing the words in the lexicon in terms of sequences of six manner of articulation labels. For the 20,000-word Webster's Pocket Dictionary, there are an average of approximately 35 words matching each broad phonetic sequence. The largest equivalence class has about 200 words, or 1% of the lexicon.

A partitioned lexicon forms a table of words corresponding to each broad class sequence. In the case of ideal data, a sequence recognized in the speech signal can be used to lookup the possible matching words in the table. This presumes that the word boundary is known, and hence applies most directly to isolated word recognition. The variability in real speech data complicates this simple access model.

Our previous research has focused on developing a lexical representation which is relatively insensitive to variability. This work is summarized in the next section. The current paper presents a method for using sequential phonetic constraints to reduce the variability in a broad phonetic classifier. To evaluate this method we implemented a simple frame-based broad phonetic classifier and tested it both with and without the sequential phonetic constraints.

LEXICAL ACCESS

The high degree of variability in speech means that a given recognized sound sequence, S, can correspond to many possible sequences, $Pos(S)$, in the lexicon. The size of $Pos(S)$ depends on both the degree of variability in the speech, and the errors introduced by the acoustic classifier.

To find all the possible words given sequence S, either the lexicon must be probed once for each sequence in $Pos(S)$, or each word must be stored according to all its possible realizations. Therefore, in order to reduce the number of word candidates corresponding to S it is necessary to minimize the size of $Pos(S)$. This can be done in two ways: (1) reduce the sensitivity of the lexical representation to variability, and (2) reduce the variability in the output of the broad phonetic classifier.

The key observation in making the lexical representation less variable is that the variability in speech is not uniform. For instance, the stressed syllables of words are less variable than the unstressed syllables. This is illustrated by the fact that deletion of phonetic segments occurs almost exclusively in unstressed syllables. Thus for two identical broad class sequences S_s and S_u, recognized from stressed and unstressed syllables respectively, the first will have fewer possible underlying sequences than the second: $Pos(S_s) < Pos(S_u)$.

In order to evaluate a representation based on stressed syllables, we compared the relative importance of stressed and unstressed syllables in partitioning a large lexicon [3]. This investigation revealed that the phonemes in stressed syllables alone provide almost as much constraint as the entire word: the size of the lexical equivalence classes is almost the same for representations using only the stressed syllables as for those using the whole word. For representations using only unstressed syllables, on the other hand, the size of the equivalence classes is two orders of magnitude larger. These results strongly suggest that the lexical representation should be based on the phonemes in stressed syllables.

The second way of minimizing the size of $Pos(S)$ is to reduce the variability in the output of the classifier. The remainder of this paper investigates how to use sequential phonetic constraints to reduce the variability in the output of a broad phonetic classifier. Since sequential phonetic constraints are implicit in the words of a given lexicon, they must be decoupled from the lexicon before they can be used in a classifier.

CH2243-4/86/0000/2259 $1.00 © 1986 IEEE

DECOUPLING THE CONSTRAINTS

This section investigates representing the sequential phonetic constraints of English explicitly in terms of allowable n-tuples of broad phonetic segments. To the extent that this representation is independent of any particular lexicon, it can be said to capture general sequential properties of English.

Sequential phonetic constraints are relatively local. For example, English has the word initial sequences "spl" and "spr", but not "spt". At a broad phonetic level (using the six manner of articulation classes vowel, nasal, liquid or glide, stop, strong fricative, and weak fricative) this rule can be characterized as

[STRONG-FRIC][STOP][LIQUID]

is allowable but

[STRONG-FRIC][STOP][STOP]

is not. The locality of such rules implies that a first or second order characterization of legal sound sequences should be sufficient for capturing sequential phonetic constraints.

N-th Order Models

Given the locality of sequential phonetic constraints, we can use the n-th order sequences (for $n = 2$) in a large corpus to construct a model of legal broad phonetic sequences. The states of the model are n-tuples of broad phonetic segments, and the transitions are single segments. A transition from state $(x_1, x_2, \ldots, x_n)$ to state $(x_2, \ldots, x_n, x_t)$ occurs on input x_t, where the x_i are broad phonetic segments.

For a broad phonetic scheme such as the one we have been using, constructing these models is relatively easy because of the small number of symbols. A third order characterization of a six symbol system, such as the manner of articulation classification used by Shipman and Zue, has only 216 possible states. For a more detailed representational scheme, with forty or fifty symbols, the number of possible states rapidly becomes intractable.

A given model can be formed by observing the broad phonetic class sequences in a particular lexicon. For example, the one-word lexicon "cast", with the phoneme string "kæst" and the broad phonetic sequence

[STOP][VOCALIC][STRONG-FRIC][STOP]

generates a second order model with three states and two transitions. However this model does not capture the legal sequences at the beginnings and ends of words. Therefore we make use of two additional classes [BEG] and [END] which mark before and after a word. Using these two additional classes, the model shown in Figure 1 is obtained for the one-word lexicon, "cast".

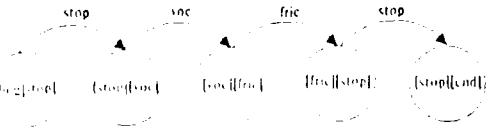


Figure 1. Second order model of a one-word lexicon.

To determine how well these models capture broad phonetic constraints independent of a given lexicon, we compared models of different lexicons. If a model of one lexicon recognizes the sequences in other lexicons, then it has captured general properties of English sound sequences rather than specific properties of the lexicon. Second and third order models of the Pocket Dictionary of 20,000 words, and Lorge and Thorndike's 3,500 most frequent English words were compared. The models use the same six manner of articulation classes as the lexicon studies. The second order model of the 3,500 word lexicon correctly recognizes 99.2% of the words in the 20,000 word lexicon. The third order model correctly recognizes 95.5% of the words. This strongly supports the fact that the models are independent of a given lexicon.

In addition to specifying all word legal n-th order broad phonetic sequences, the networks can be used to encode the likelihood of occurrence for each sequence. This is done by augmenting the arcs of the network with transition probabilities. Using a lexicon with word frequencies, the likelihood of a given transition is proportional to the frequency of the words in which it occurs. Our lexicons all have word frequency information from the Brown Corpus of written English [4]. With the addition of transition probabilities, the networks form n-th order Markov models of the sequential phonetic constraints.

APPLYING THE CONSTRAINTS

The goal of incorporating sequential phonetic constraints into a classifier is to reduce the variability in the classifier's output. To evaluate the effectiveness of the models developed in the previous section we implemented a simple broad phonetic classifier. Given the output of the classifier, a Markov model is used to determine the best broad phonetic sequence consistent with that model. Figure 2 diagrams the relation between the classifier and the model. In effect the sequential phonetic constraints of the model are used to "correct" the output of the classifier. Comparing the best sequences for different models serves as a paradigm for evaluating the power of the models.

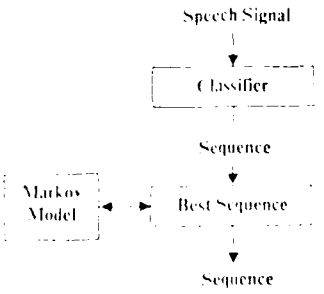


Figure 2. Paradigm for comparing different Markov models of sequential phonetic constraints.

The n-th order Markov model of a given lexicon captures sequential phonetic constraints in terms of broad phonetic segments. Since the classifier performs frame-by-frame classification of the speech signal, the segment-based networks of the previous section must be converted to frame-based networks.

Each arc in a segment-based network corresponds to a given broad phonetic segment, as illustrated in Figure 1. To convert this network into a frame-based network, each arc is replaced by a recognizer for its corresponding segment. This segment recognizer models a segment as one or more successive frames of the same type, as shown in Figure 3. The self-transition probability, p_t, is obtained by observing the durations of each broad

Figure 3 Frame-by-frame model of a broad phonetic segment

phonetic segment in a training set. The small number of broad phonetic segments makes robust training possible with relatively little data.

The same segment recognizer is used regardless of what states a transition occurs between. One obvious extension is to determine the degree to which broad phonetic context influences the acoustic realization of broad phonetic segments. If context predicts the acoustic realization, then different segment recognizers can be used in different contexts.

The network formed by replacing each segment arc with its corresponding frame-based recognizer forms a Markov model of frame sequences. This model captures both the n-th order segment information in the lexicon and the duration information in the training set.

Finding the Best Frame Sequence

Given a sequence of frames generated by the classifier, we wish to determine the best frame sequence given that input and an n-th order frame-based Markov model. This can be done using the forward-backward algorithm [5], and defining the best frame sequence to be the highest probability sequence given the model and the input sequence.

The highest probability sequence is simply the highest probability state at each time. The highest probability state at a given time is found by using the forward-backward algorithm to compute $Pr(s_t = q_i \mid \mathcal{O})$, the probability of being in state q_i at time t given the observed sequence $\mathcal{O}$. This computation is done using the conditional probability

$$Pr(s_t = q_i \mid \mathcal{O}) = \frac{Pr(\mathcal{O} \text{ and } s_t = q_i)}{Pr(\mathcal{O})} \qquad (1)$$

and the relation

$$Pr(\mathcal{O} \text{ and } s_t = q_i) = \alpha_t(i)\beta_t(i)$$

where

$$\alpha_t(i) = Pr(O_1 \cdots O_t \text{ and } s_t = q_i)$$
$$\beta_t(i) = Pr(O_{t+1} \cdots O_T \mid s_t = q_i)$$

are the *forward* and *backward* probabilities, respectively. These probabilities and the denominator of (1) can be computed efficiently (in $O(n^2)$ time for an n state model) using recursion formulas.

The next section describes a classifier which vector quantizes the output of three bandpass filters into eight VQ codewords [6]. These codeword sequences are input to the forward-backward algorithm along with the likelihood of each broad phonetic class given a particular codeword.

THE CLASSIFIER

The input to the classifier is a crude spectral shape in the form of three bandpass filtered energies in the ranges 0-1000 Hz, 1000-2500 Hz and 2500-8000 Hz, computed every 10 milliseconds. Each energy value is computed using a 20 millisecond hamming window. The three energy values at each frame are then vector quantized into one of eight VQ codewords. This VQ codeword sequence is used as input to the forward-backward algorithm.

The training procedure involves three stages. The same set of training utterances is used for all three stages. The first stage estimates the self-transition probabilities, p_t, for the segment recognizers used to construct the frame-based network. These probabilities are determined using the durations of the hand-labeled data segments.

The second stage forms the vector quantization codebook. The three bandpass energy values for each frame of the training data are input to a k-means clustering procedure, using a Euclidean distance metric. The VQ codewords are the centroids of the resulting clusters.

The third stage estimates the likelihood of the broad phonetic classes given each VQ codeword. These probabilities are estimated from hand labeled data and the output of the vector quantizer. Given the small number of broad classes, these probabilities can be estimated from relatively little data.

RESULTS

The classifier was evaluated using both four and five broad phonetic classes. The four classes are: vocalic (VOC), voiced closure (VCL), noise (NZ) and silence (SIL). The five classes differ only in the replacement of the single class NZ by the two classes fricative (FRC) and burst (BST).

The 1892-word lexicon from the Harvard List sentences was used to form the Markov models of broad phonetic sequences. This lexicon is similar in complexity to the Lorge-Thorndike and Pocket dictionaries described above. A second order model of the Harvard lexicon using the five broad classes recognizes all the words in the 3,500 and 20,000 word dictionaries. A third order model recognizes 92% and 89% of the two lexicons, respectively.

Performance was measured using both frame and segment-based statistics. The frame-by-frame performance compares the best frame sequence output by the forward-backward algorithm against the hand label for each frame. This yields both a confusion matrix and an overall percent-correct figure for the frame-by-frame classification.

The segment-based performance is computed by chaining together successive frames with the same label. The resulting "segments" are then compared against the hand labeled segments by computing a best match between the two segment sequences. The match is constrained such that segments must occur in time in order to be matched. This provides a more accurate (and more conservative) performance measure than a best string match.

The system was tested separately for each of thirty male speakers. Each speaker said ten sentences from the Harvard List. For each test, the training set consisted of the remaining twenty-nine speakers. This procedure was used to establish the speaker

independence of the system. Table 1 shows the frame-by-frame performance averaged across all thirty speakers for both four and five broad phonetic classes. Three different sequential phonetic models were evaluated. The first row shows the results for the "0-th order" models, which incorporate durational information but no sequential phonetic constraints. The second row shows the results for the first order models. Since the results for the second order models are almost the same, they are not presented.

Order	4 Classes	5 Classes
Zero	67.3%	67.1%
First	75.1%	71.8%

Table 1 Average correct frame-by-frame classification for thirty speakers. System was trained and tested separately for each speaker; the test speaker was not used for training.

These results demonstrate that adding sequential phonetic constraints increases the frame-by-frame recognition performance over the zero-order model. Using more sophisticated segment recognizers to convert the segment-based to frame-based models could further increase the frame-by-frame performance.

Table 2 shows the segment-based performance. The percentage of segments correctly classified is reported, along with the segment insertion rate in parentheses. The sequential phonetic constraints have a substantial effect in reducing the segment insertion rate, without greatly decreasing the percentage of the segments correctly recognized. Recall that these results are relatively conservative because automatic and hand labeled segments must overlap in time in order to be considered a correct match.

Order	4 Classes	5 Classes
Zero	81.5% (63.7%)	83.0% (78.4%)
First	74.1% (11.0%)	75.1% (11.8%)

Table 2 Average segment-based correct classification for thirty speakers. The segment insertion rate is in parentheses.

The segment insertions for the first order models are highly regular. For instance 66% of the insertions in the 5-class condition are VCL in a FRC VOC context. Thus additional processing of the segments should be able to substantially reduce the 11% insertion rate. Without the sequential constraints the errors show no such regular patterns, and hence further processing is not likely to reduce the error rate.

Syllable Stress Affects Classifier Performance

The classifier's segment deletion rate is higher in the unstressed syllables than in the stressed syllables. In making this comparison, only mono-syllabic words were considered (84% of the words in the utterances are mono-syllabic). A word was called unstressed if the nuclear vowel was reduced transcribed as a schwa and otherwise was called stressed. For the unstressed words the deletion rate was 23.7% whereas for the stressed words it was only 15.9%. This result adds further support to the earlier observation that the stressed syllables are important in hypothesizing words.

SUMMARY

While lexicon studies demonstrate the power of broad phonetic constraints for differentiating words from one another, they do not suggest how such constraints can be directly exploited in recognition. This paper has presented a method for decoupling sequential phonetic constraints from a given lexicon, by representing allowable broad phonetic sequences in terms of n-th order Markov models. Tests of a simple frame-based broad phonetic classifier on 300 sentences from 30 speakers demonstrate that these models can be used to increase the performance of a broad phonetic recognizer.

Acknowledgments

This work was done at the Speech Communications Group and the Artificial Intelligence Laboratory at MIT. Support was received in part from the Office of Naval Research under contracts N0014-82-K-0727 to the Speech Group, and N0014-80-C-0505 to the AI Lab. The VQ and network algorithms were implemented with Gary Kopec at Schlumberger Palo Alto Research.

REFERENCES

[1] D.P. Huttenlocher and V.W. Zue, "A model of lexical access from partial phonetic information", *Proc. ICASSP*, 1984.

[2] D.W. Shipman and V.W. Zue, "Properties of large lexicons: implications for advanced isolated word recognition systems", *Proc. ICASSP*, 1982.

[3] D.P. Huttenlocher, "Sequential phonetic constraints in recognizing spoken words", MIT Artificial Intelligence Laboratory, Memo No. 867, 1985.

[4] H. Kucera and W.N. Francis, *Computational Analysis of Present-Day American English*, Brown Univ. Press, 1967.

[5] S. Levinson, L. Rabiner and M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", *Bell System Technical Journal*, **62(4)**, 1983.

[6] R. Gray "Vector quantization", *IEEE ASSP Mag*., 1984

From the Proceedings of ICASSP 86, the IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, held in Tokyo, Japan, April 8-11, 1986.

VISUAL CHARACTERIZATION OF SPEECH SPECTROGRAMS*

Hong C. Leung and Victor W. Zue

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper describes a system that applies vision techniques to extract acoustic patterns in the speech spectrogram. By processing a spectrographic image through a set of edge detectors and combining their outputs, the system obtains two-dimensional objects that characterize the formant patterns and general spectral properties for vowels and consonants. As a validation of the approach, a limited vowel recognition experiment was performed on the "object" spectrograms. Preliminary results show that this processing technique retains relevant acoustic information necessary to identify the underlying phonetic representation.

INTRODUCTION

For the past four decades, the prevailing form for displaying speech has been the spectrogram, a three-dimensional time-frequency-intensity representation of the signal. The spectrogram provides a visual display of the relevant temporal and spectral characteristics of the acoustic signal. It has been an invaluable tool in the development of our understanding of the acoustic properties of speech sounds.

Recently, the spectrographic display took on added significance as it was demonstrated that the underlying phonetic representation of an unknown utterance can be extracted almost entirely from a visual examination of the speech spectrogram [2], [3], [9]. In these experiments, a trained spectrogram reader correctly identified the phonetic segments with 80% to 90% accuracy, depending on the experimental conditions and the scoring procedures. The reader's performance, measured in terms of accuracy and rank-order statistics, was considerably better than that of the phonetic front-ends of available speech recognition systems. These experiments stirred renewed interest in acoustic-phonetic approaches to speech recognition, and supported the speculation that better front-ends may be constructed if we can learn the phonetic decoding procedure used by human experts.

Protocol analysis of spectrogram reading reveals that the decoding process calls for the recognition and integration of a myriad of acoustic patterns. In order to develop a system that utilizes such knowledge, one must first be able to extract these acoustic patterns.

This paper is concerned with the visual characterization of speech spectrograms. Our aim is to capture the essential acoustic patterns of a spectrogram so that these abstracted patterns may be used to characterize and recognize different speech sounds. Traditional descriptions of acoustic-phonetic events based on formant frequencies are often inadequate because the formants cannot always be resolved reliably. Thus visual characterizations may provide an alternative, and perhaps more effective, description.

Processing the spectrogram as a three-dimensional image has a number of important advantages. First, one can better capture the time-frequency dependency of the speech signal by treating the time and frequency dimensions simultaneously. Second, we can liberally borrow from techniques developed through many years of successful vision research. Third, characterizing a spectrogram is a highly constrained vision task. The three dimensions of the spectrogram correspond to physically meaningful quantities, namely, time, frequency, and amplitude. The patterns on the spectrogram are also limited by the nature of the speech production mechanism and the restricted sound patterns of a language.

SYSTEM DESCRIPTION

Our approach to visual characterization of speech spectrograms is to treat the acoustic patterns as visual objects. These objects are obtained by applying edge detection to the spectrographic image, producing an "edge map" as output. The edge map includes explicit information about the position, the orientation, and the relative strength of edges. These edge elements are grouped into closed geometrical contours. The remainder of this section describes the system in greater detail, focusing on the vowel-like sounds. Obstruent sounds have visual patterns that are quite different from those of vowel-like sounds. Their treatment will be described near the end of this section.

Edge Detection

The system obtains a narrow-band spectrographic representation by computing a short-time spectrum once every 5 ms with a 25.6 ms window. The vowel-like regions of the image, determined through a broad phonetic classifier [5], are then processed through two-dimensional directional edge detectors of different scales. The cross-section in the frequency dimension is the second derivative of a Gaussian, and the cross-section in the time dimension is a Gaussian. The directional Gaussian edge detector has been shown by Canny [1] to have many useful properties such as robustness against detection errors, good localization to

*This research was supported by DARPA under contract N00014-82-K-0727, monitored through the Office of Naval Research.

CH2243-4/86/0000/2751 $1.00 © 1986 IEEE

true edges, and dimensional separability. Thus this operator smooths the spectrogram in the time dimension and also detects edges that are approximately orthogonal to the frequency dimension. Zero-crossings of the filtered output correspond to edges in the original spectrogram. Another advantage of using Gaussian detectors is that the zero-crossings do not disappear as the scale[1] decreases [7], [8]. This is an important property for combining outputs from different scales.

One potential problem of using a directional operator is that its performance might degrade if the formants are not quite horizontal. Multiple directional operators oriented at different angles might, therefore, be needed. However, due to the sluggishness of articulatory movements, formant frequencies cannot change very quickly. Preliminary results show that if the Gaussian cross-section in the time dimension is made small enough (on the order of 1 pixel), the edge detector can pick up fast formant movement.

Combining Multiple Scales

Parts (a) and (c) of Figure 1 show the narrow-band and wide-band spectrograms, respectively, for the nonsense word, "boyt", spoken by a female speaker. Parts (b), (c), and (d) show the results of filtering the narrow-band spectrogram with the directional edge detectors of different scales. The plots correspond to a σ of 4, 3, and 2 pixels, with sigma decreasing from left to right in the figure. The output with the largest scale is the most robust but has the least resolution, whereas the one with the smallest scale has the best resolution but also has many extraneous edges. In order to achieve robustness and good resolution simultaneously, these outputs must be systematically combined.

We have chosen to combine the outputs by performing a coarse-to-fine tracking in a way similar to scale-space filtering proposed by Witkin [7]. This approach has the advantage of managing the ambiguity of scale in an organized and natural way. Since zero-crossings do not disappear as the scale decreases, the coarse-to-fine tracking works properly.

Figure 1(f) illustrates the result after combining edges from the different scales. It can be seen that the result has good resolution and is robust.

Applying Speech Knowledge

While coarse-to-fine tracking solves the problem of localizing large-scale events, it does not solve the multi-scale integration problem. Which of the edges found by the small-scale operators are robust, and which edges are due to noise? There are a number of ways to determine which edges are valid. One measure is to examine the amount of intensity change. The amplitude of the output of the first derivative Gaussian, and the slope of the zero-crossings of the second derivative Gaussian, are good indicators of the amount of intensity change. However, some form of thresholding is needed, which may lead to gross error.

We have chosen, instead, to apply specific speech knowledge to select the edges. We first apply a bandwidth constraint. For some vowels, formants can be quite close to each other. Sometimes they are so close together that it is impossible to separate them by eye. Spectrogram readers are able to tell that there are two formants because of the bandwidth. Thus after the coarse-to-fine tracking is performed, regions with significantly large bandwidths are suspected of having more than one formant. In these cases, edges from the smaller operator outputs can be included if the bandwidths after the insertion of the additional edges are still reasonable. This heuristic is quite robust in the vowel regions. To avoid including spurious edges, however, the original bandwidth needs to be quite large so as to trigger insertion of edges. This means that some of the good edges from the smaller-scale detectors are inadvertently omitted. In order to locate these edges, more elaborate procedures are needed.

For some vowels, the formants are quite close to each other for some duration, but gradually separate and finally split apart. After the formants split, edges can be detected quite reliably. These edges can then be used as anchor points to find edges when the two formants approach each other. As we have seen in Figure 1(f), F1 and F2 begin to split apart at approximately the *midpoint* of the vowel. This kind of split provides strong evidence that more edges should lie to the left of this point. These subtle edges are located by the following "digging" procedure. Starting from this point, edges to the left are examined. If these edges satisfy a continuity requirement, they are considered "good" edges. Building upon the extensions, edges further to the left are then examined. This process repeats until no more edges are found or until the continuity constraint is violated. Figure 1(g) shows the result after the "digging" operation. In this example, the operation has dug through the entire region and correctly located the first two formants of the vowel. (Note also that objects with average frequency above 3.5KHz have been discarded, since they do not contribute to the phonetic identity of vowels.)

The scale-space filtering, augmented with the above two procedures, is quite robust in finding formant edges in the vowel regions. At relatively high frequencies, the detected edges usually correspond to edges of the formant frequencies. However, there is very often an energy concentration below 300 Hz due to F0. When F1 is low, this small energy concentration is masked by F1. But when F1 is higher in frequency, this energy concentration becomes more and more noticeable. Trained spectrogram readers are very good at ignoring it. We are not yet sure how to deal with these shallow edges in the system. At this moment, we have chosen to ignore edge contours with average frequency less than 300 Hz if there is another edge contour with average frequency below 800 Hz. This condition ensures that the ignored contour does not correspond to F1.

Processing of Obstruent Regions

Obstruents are characterized by their general spectral distributions rather than any specific formant patterns. As a result, the processing for the obstruent regions is considerably different from that of sonorant regions. The obstruent regions are again determined by the broad phonetic classifier. A very coarse edge detector is applied to the *wide-band* spectral slices, computed with a 6.7 ms window. The objects are obtained from the edge map with no further processing.

Figure 1(h) shows the final result for the word "boyt," including both the vowel-like and obstruent-like regions. Comparing this figure with the original spectrogram, we see that rele-

[1] The scale is a measure of the width of an edge detector. For a Gaussian detector, the scale corresponds to the standard deviation, σ.

vant features in the original spectrogram have been captured in the objects. As a more elaborate example, Figure 2(b) shows the objects obtained from a continuous sentence spoken by a male speaker. For comparison, the corresponding wide-band spectrogram is shown in Figure 2(a). If the extracted objects indeed capture the important information in the spectrogram, then they can be used as a mask to filter out irrelevant acoustic information, as shown in Figure 2(c). We see that important acoustic information in this utterance, such as the formant transitions in vowel regions and the shift in spectral energy distributions in obstruent regions, has been accurately retained after processing.

RECOGNITION EXPERIMENTS

The examples shown in Figures 1 and 2, and informal "object-reading" experiments performed by spectrogram-reading experts, suggest that the procedure described in the previous section is potentially useful in extracting important acoustic features from the spectrograms. The extracted patterns can, for example, provide the necessary information for the development of a knowledge-based system for phonetic recognition [10]. Alternatively, one can build up an inventory of these patterns in order to characterize and recognize speech sounds directly, using a variety of visual object recognition algorithms [6]. Before we start to utilize these objects in either of the two tasks, however, we must first make sure that these processed visual patterns indeed retain the necessary information for the recognition of the underlying phonetic segments,

As a step in this direction, we performed a small vowel recognition experiment. The task involves the recognition of 14 vowels, /i, ɪ, e, ɛ, æ, ɑ, ɔ, ʌ, o, u, ɝ, ay, ɔy, aw/, spoken in the /b/-vowel-/t/ environment by 8 male speakers. Due to the limited amount of available data, the recognition was performed using a rotational procedure; in each trial the system was trained on the data from seven speakers and tested on the remaining one. For each vowel, the recognizer chose from the seven training samples the one with the smallest intra-sample distance as the reference template. A dynamic time warping algorithm [4], with appropriate local path constraints, was used to compensate for differences in duration between the test and reference patterns. No attempt was made for normalizing the frequency scale to account for inter-speaker differences.

The objects determined by our processing system do not retain amplitude information which is often useful in characterizing speech sounds. Therefore, we created from the objects a cartoonized spectrum for each time frame. Regions inside the objects were replaced by a constant value that is equal to the average value of the corresponding regions in the original spectrum, whereas regions outside were set to zero. The cartoonized spectrum was then smoothed with a Gaussian window. Parts (a), (b), and (c) of Figure 3 illustrate, respectively, a vowel spectrum (superimposed by an LPC spectrum), the cartoonized spectrum derived from the edges, and the smoothed spectrum used for recognition. A Euclidean distance was used to measure similarities between spectra. For comparison, we also implemented an LPC-based system using the Itakura's distance metric [4].

The results of our vowel recognition experiments, based on the 112 vowel tokens from eight speakers, show that the smoothed spectra can be used to identify the vowels with an 83% first-choice accuracy. The correct vowel is within the top two choices 94% of the time. This result compares favorably to that using the LPC/Itakura-Distance method. While it is premature to base our conclusion on such a restricted corpus, we are nevertheless encouraged by the results. It appears that, for this data set at least, our processing system did not remove acoustic information that is necessary for vowel identification.

SUMMARY

In summary, we developed an algorithm for the extraction of visual objects from speech spectrograms. Results from a limited vowel recognition experiment suggest that the processing technique retains acoustic information that is useful for phonetic distinction.

In the future, we plan to evaluate this system more extensively, and to investigate the feasibility of using the objects for phonetic recognition.

REFERENCES

[1] Canny, J.F., "Finding Edges and Lines in Images," MIT-TR-720, MIT.

[2] Cole, R.A., Rudnicky, A.I., Zue, V.W., and Reddy, D.R, "Speech as Patterns on Paper," in *Perception and Production of Fluent Speech*, R.A. Cole, ed., Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980, pp. 3-50.

[3] Cole, R.A. and Zue, V.W., "Speech as Eyes See It," in *Attention and Performance VIII*, R.S. Nickerson, ed. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980, pp. 475-494.

[4] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 1, pp. 67-72, Feb. 1975.

[5] Leung, H.C. and Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech" *IEEE Conference Proceedings, ICASSP*, San Diego, CA, 1984, paper 2.9.

[6] Marr, D., *Vision*, W.H. Freeman & Co., San Francisco, 1982.

[7] Witkin, A.P., "Scale-Space Filtering," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1019-1022, 1983.

[8] Yuille, A.L. and Poggio, T., "Scaling Theorems for Zero-crossings," AI Memo 722, MIT.

[9] Zue, V.W. and Cole, R.A., "Experiments on Spectrogram Reading," *IEEE Conference Proceedings, ICASSP*, Washington D.C., 1979, pp. 116-119.

[10] Zue, V.W. and Lamel, L.F., "An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition," *IEEE Conference Proceedings, ICASSP*, Tokyo, Japan, 1986, paper 23.2.

51. 1. 3

From the Proceedings of ICASSP 86, the IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, held in Tokyo, Japan, April 8-11, 1986.

AN EXPERT SPECTROGRAM READER: A KNOWLEDGE-BASED APPROACH TO SPEECH RECOGNITION*

Victor W. Zue and Lori F. Lamel

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

Human experts can determine the phonetic identity of unknown utterances from a visual examination of the spectrogram with performance better than available computer systems. The spectrogram-reading process involves the use of multiple sources of knowledge, including articulatory movements, acoustic phonetics, phonotactics, and linguistics. In addition, the experts' performance can be attributed to their ability to deal with partial and/or conflicting information, as well as multiple cues.

This paper investigates the feasibility of constructing a knowledge-based system that mimics the process of spectrogram reading by humans. In a task of identifying stop consonants extracted from continuous speech, the system achieved performance that is comparable to that of the experts.

INTRODUCTION

Over the past four decades the spectrogram, a three-dimensional time-frequency-intensity representation of the signal, has been the single most widely used form of display for speech. Part of its popularity stems from the fact that it is relatively easy to produce, and it provides a visual display of the relevant temporal and spectral characteristics of the acoustic signal. It has been an invaluable tool in the development of our understanding of the acoustic properties of speech sounds.

Recently, a series of experiments by Zue and his colleagues demonstrated that the underlying phonetic representation of an unknown utterance can be recovered almost entirely from a visual examination of the speech spectrogram [1], [2], [3]. In their experiments, a trained spectrogram reader correctly identified the phonetic segments with 80% to 90% accuracy, depending on the experimental conditions and the scoring procedures.

While the spectrogram-reading experiments were intended to illustrate the richness of phonetic information in the speech signal, the results are relevant to automatic speech recognition in several respects. First, they demonstrate that a great deal of phonetic information can be derived from the acoustic signal alone. The reader's performance, measured in terms of accuracy and rank-order statistics, was considerably better than that of the phonetic front-ends of available speech recognition systems. The experiments thus provide an "existence proof" that high-performance phonetic recognition is attainable. Second, spectrogram reading is based on the recognition and integration of a myriad of acoustic cues. Some of these cues are relatively easy to identify, while others are not meaningful until the relevant context has been established. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. The discovery of the acoustic cues and, more importantly, of the control strategies for utilizing these cues are the keys to high-performance phonetic recognition. Finally, protocol analysis of the process of spectrogram reading reveals that the decoding process often involves the use of explicit rules. Thus the knowledge used in spectrogram reading is potentially transferable to others, both humans and machines.

Our experience with spectrogram reading suggests that the reasoning process can be naturally expressed as a series of production (or *if-then*) rules, where the preconditions and conclusions may be phonetic features or acoustic events. Since the acoustic-phonetic encoding is highly context-dependent and redundant, we must be able to entertain multiple hypotheses and to check for consistency. Acoustic features are often expressed in a qualitative manner and described as being present/absent, and having values such as high/mid/low, or weak/strong. Thus in order to have the computer mimic the performance of spectrogram readers, we need a system that can deal with qualitative measures in a meaningful way.

In this paper, we report preliminary results of our attempt to incorporate our knowledge about the spectrogram-reading process in a knowledge-based system that mimics the process of feature identification and logical deduction used by experts. The knowledge base explicitly represents the expert's knowledge in a way that is easy to understand, modify, and update. Our research direction is very similar to the efforts by Johanssen et al. [4] and Johnson et al. [5].

TASK DEFINITION

The process of spectrogram reading involves extracting relevant acoustic features and combining these features using rules that relate the underlying phonetic forms to their acoustic manifestations. Our task investigates the feasibility of developing a computer system that mimics such a process.

In order to keep the project manageable, we made some important design restrictions. First, we decided to focus on the acquisition and formalization of the knowledge base, rather than the development of an expert system itself. As a result, our initial effort makes use of an available *Mycin*-based[6], backward-chaining[1]system. Our investigation thus far has revealed that this particular expert system may not be the most appropriate.

*This research was supported by DARPA under contract N00014-82-K-0727, monitored through the Office of Naval Research.

CH2243-4/86/0000-1197 $1.00 © 1986 IEEE

Nevertheless, it has provided us with a convenient mechanism to acquire and formalize our knowledge, while freeing us from the need to delve into a very difficult research area.

Second, we bypass the problem of automatic extraction of acoustic features. Many of the acoustic features used during spectrogram reading are readily extracted by the human visual system, but are very difficult to extract automatically by computer. For example, there does not yet exist a formant tracker that can determine formant frequencies reliably, especially in regions where the direction and the extent of formant transitions provide important information about the place of articulation for consonants. Thus, while the measurements were made automatically whenever possible, the acoustic features were verified by the experimenter before being entered into the database. Recent work by Leung and Zue [7] attempts to locate two-dimensional objects directly from the spectrogram. Their work on visual object recognition may eventually play a role in the feature extraction part of our system.

Finally, we selected the task of identifying stop consonants both as singletons and in clusters, since the cues for stop consonants are complex, interrelated, and easily modified by phonetic context. This paper reports on the identification of word-initial singleton stop consonants that appear between two vowels. Stops have been extensively studied and recognition results are available for comparison.

SYSTEM DESCRIPTION

The development of our knowledge-based system for spectrogram reading is divided into two parts. First we select a set of acoustic features that are important for phonetic decoding, and outline the procedures for their extraction. Then we develop rules that operate on these acoustic features to deduce the underlying phonetic form. This latter task involves both the formalization of our knowledge with respect to the terminology and descriptions, and the actual statements of the acoustic-to-phonetic mapping. These two aspects of the system are described next.

Making the Measurements

Feature Selection The acoustic features useful for specifying a given phonetic contrast were initially determined by combing the acoustic-phonetic literature and by observing spectrogram reading sessions conducted by experts. Next, several hundred spectrograms containing stop consonants were annotated by experts and studied to verify the usefulness of these cues and to suggest supplementary measurements. For our current task of stop identification, we obtained acoustic features that describe the release burst, the closure interval, and the surrounding contexts. These features include the voice onset time (VOT), the location and the strength of the burst, and the formant transitions preceding closure and following release. Our system currently utilizes 26 acoustic features.

Feature Extraction As stated earlier, at this moment we are not concerned with the automatic extraction of the acoustic features. Instead, we assume that the measurements of the acoustic features are made with no error, and as a result, system performance can be assessed in relation to the adequacy of the acoustic features, the rules, and the control strategy.

[1]Many search problems can be treated as finding a path to a goal state from some initial position. When the search proceeds from the initial state toward the goal state, it is said to be a forward chaining system. In contrast, when the search starts at the goal state and works back toward the initial state, then it is said to be backward chaining.

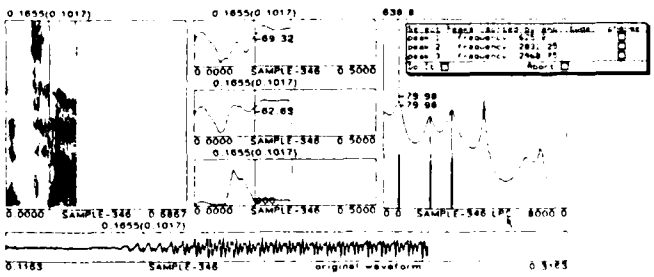


Figure 1: A display of the interactive measurement system.

Some of the acoustic features can be measured reliably without human intervention. For example, the system can automatically determine whether the following vowel is rounded from the phonetic transcription. Some other measurements, such as whether the stop release is pencil-thin, are qualitative in nature and must be provided by the expert. Most of the measurements, however, can be made automatically, subject to verification by the expert. For example, although the time location of the burst is a measurement first made by the computer, verification is necessary partly because the measurement is inherently error prone, and partly because other measurements depend on accurate burst location.

To facilitate the measurement of the acoustic features by hand, we have developed a semi-automatic system that makes many of the measurements automatically based on a time-aligned phonetic transcription [8]. In making measurements, the expert has available displays of the spectrogram, the speech waveform, the short-time spectra, and energies in selected frequency bands. The system goes through a checklist of acoustic features, making the measurements and querying the expert to verify or modify them. An example of the display used by the expert to make the measurements is shown in Figure 1. In this example, the system determined the first three formants at the onset of the following vowel without error. The formant frequencies are marked by a short vertical line, with associated numerical values, in the short-time spectrum window at the upper right-hand corner of the display.

Each sample in the database has an associated list of feature values that are mostly numerical. These values are used to develop rules and to test the knowledge-based system.

Formalising the Knowledge

Not much is known about how experts approach the spectrogram-reading problem. The general strategy of expert spectrogram readers is to make some preliminary proposal separating the segments into broad phonetic classes. The candidate set is then refined by incorporating detailed acoustic cues to rule out unlikely hypotheses. In our attempt to capture this complicated problem-solving procedure, we employ several general principles. First, multiple hypotheses based on diverse acoustic evidence must be entertained. Second, the presence of a cue may be useful, but its absence need not be harmful. Third, very strong evidence of one kind may preclude competing hypotheses. An example utilizing these principles is shown in Figure 2. The place of articulation of the stop consonant in the right-hand panel can be readily identified as VELAR by the compact,

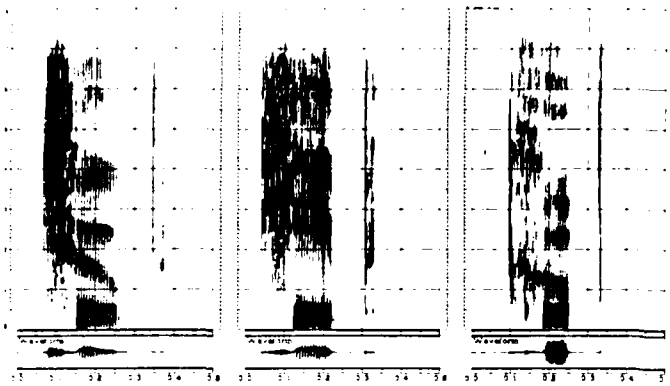


Figure 2: Spectrograms of /t/ and /k/ preceding different vowels.

low-frequency burst. No other information is necessary. On the other hand, the bursts for the other two stops are very similar; both are rich in high-frequency energy. Only after the vowel context is known can one infer that the first stop is ALVEOLAR (in a rounded environment) and the second stop is VELAR (in a fronted environment).

In our system, phonemes are represented as a bundle of distinctive features [9]. Thus, for example, the stop /t/, has the features: STOP, VOICELESS, ALVEOLAR. A stop is identified when there is strong evidence for the presence of its distinctive features. Our system uses three stages to identify stops. First, the phonemes are mapped into a set of distinctive features. Next, the numerical values of the acoustic features are mapped into a set of qualitative descriptions, such as high/low and strong/weak. Finally, a set of relatively independent rules deduce each distinctive feature from the qualitative descriptions.

Structure of the Rules There are several types of rules in our system, each dealing with a particular transformation of the data. First, there are rules that define the relationship between a phoneme and its distinctive feature values. For example, the stop /t/ is defined by the following rule:

> **If** the voicing of the stop is *voiceless*,
> **and** the place of articulation of the stop is *alveolar*,
> **then** the identity of the stop is /t/.

All of the stops are defined in the same manner. Thus we have converted the problem of deducing the identity of a stop to one of determining its voicing and place characteristics.

When experts read spectrograms they use their visual system to extract features in the image. Then, using a wealth of knowledge, they combine these features to form phonetic hypotheses. Experts use qualitative descriptions, such as a second formant that is low, mid, or high, but rarely specify numeric values. Although they have an intuitive sense of what these terms mean, experts may have difficulty quantifying them reliably.

In order to simulate this process, a set of rules has been developed to map the numerical values of the acoustic measurements into qualitative descriptions. The mapping ranges have all been hand-selected from histograms. Generally the qualitative descriptions are associated with disjoint numerical regions. Measurements that fall between regions are associated with *both* labels, each with a lower confidence factor.

The last set of rules deduces the distinctive features from the acoustic descriptions. Two examples of rules that deduce the feature VOICING are shown below:

> **If** the VOT is *short*,
> **and** the following vowel is not a *schwa*,
> **then** the stop is *voiced*.

> **If** there is prevoicing during closure,
> **then** the stop is *voiced*.

The second example reflects the asymmetry of some of the acoustic cues; in this case the presence of prevoicing is a good indicator for a voiced stop whereas the absence of this cue does not necessarily rule out a voiced stop. Note also that the strength of a rule's conclusions depends upon the belief in the preconditions. If one is uncertain about the acoustic measurements, multiple rules can be fired, each with a lower confidence factor.

Control Strategy *Mycin* uses a very simple goal-directed control strategy. It sets off to determine the identity of the stop, and in the process needs to deduce its voicing and place characteristics. In each case, the system will exhaustively fire all the pertinent rules. We are able to affect the control strategy somewhat by including preconditions that inhibit certain rules from firing. For example, if the stop release is very weak, one should not pay attention to the frequency location of the burst, as it will be unreliable. As another example, the formant transitions for voiced stops are measured after voicing onset. However, for voiceless stops, the same measurements are made during aspiration, since the transitions are already completed by voicing onset.

EXPERIMENTAL RESULTS

To test the effectiveness of our system we performed a stop identification experiment in which the stops are known to be word-initial and to appear between two vowels. We greatly reduced the complexity of the problem by restricting our information to the segment to be identified and its immediate neighbors. In making the measurements, the system was provided with knowledge of the vowel contexts and with time points that roughly correspond to the points of closure, release, and voicing onset. Refined time-points and other measurements were determined using the interactive system described earlier.

Data Description

Two hundred intervocalic stops were randomly selected from a database of 1,000 sentences spoken by 100 speakers, 50 male and 50 female. One hundred tokens were used for system training, and 100 for system testing. The stops for the training and test sets were obtained from 64 speakers; 45 appear in both data sets. There was no restriction on the vowels; in fact, some of the stops preceded a schwa. In order to compare the system's performance to human performance, spectrograms of the training and testing samples were read by five experts.

System training involves selecting the acoustic features, setting the thresholds for the mapping functions, and formulating the rules. Rule development is an iterative process; an initial set of rules is proposed and tested on a subset of training samples. By examining the output of the system, the experimenter refines the rules and tests them on other training samples. The process continues until the system behavior is judged to be satisfactory.

condition		first choice accuracy	top 2 choice accuracy
training	human(2)	90	92
	system	88	95
testing	human(3)	92	96
	system	84	92

Table 1: Comparison of human and system identification performance

Performance Evaluation and Discussion

Table 1 summarizes the results of our experiments. For the training data, the system's performance is comparable to that of the experts. The performance of the system degraded by 4% when it was confronted with new data, whereas the experts' performance on the test data remains high. We attribute the degradation of performance from training to test data primarily to the "lack of experience" of the system; it has not yet learned all the acoustic features and rules used by the experts. Most of the errors are not due to new speakers, and there is no obvious male/female bias.

Table 2 displays the confusion matrix on the system's first choice identification for the test data. All but one of the errors are in identifying the place of articulation. Ten of the 16 errors involve the VELAR place of articulation. Examination of the spectrograms reveals that most of the errors made by the system are judged to be reasonable by experts. For example, /t/-/k/ confusion usually occurs when the /t/ is rounded, /k/-/t/ confusion when the /k/ is fronted, and /k/-/p/ confusion when the /k/ is back and has a weak release.

We are encouraged by the initial performance results of our system. Although the system did not perform as well as human experts, our results are comparable to stop recognition results reported in the literature on similar tasks. While stops have been extensively studied, most recognition experiments reported have been on word-initial stops in isolated words and/or pre-stressed position. The recognition task closest to our own was reported by Demichelis et al [10]. Using acoustic features that were combined with fuzzy logic and rules, they achieved recognition rates of 90 92% for stops in continuous speech.

SUMMARY

We believe that we are making headway in our efforts to capture the knowledge used by experts in the spectrogram-reading task, and to encode that knowledge into features and rules. While the rule set is still incomplete, we feel that the rules express our knowledge succinctly. As stated earlier, rule development is an iterative and interactive process. Each iteration improves our knowledge and understanding, which is then reflected in the system design and performance. As more and more data is used for training, statistical techniques can be employed to arrive at a more accurate measurement-to-description mapping.

While the performance of the system can be improved, the current implementation does not accurately model the problem-solving procedure used by human experts. This is partly due to limitations imposed by the structure of the *Mycin*-based expert system that we are using. The goal-directed, backward-chaining inferencing of *Mycin* does not enable the system to evaluate multiple hypothesis at any given time. As a practical matter this makes the system harder to use and debug. In contrast, experts tend to do forward induction, and to keep a set of possible candidates. In the future, we plan to implement our rules in a forward chaining system that better models expert behavior. We also intend to evaluate the system more extensively, and to increase the complexity of the task by extending the recognition to include impostors and stops in clusters.

stop identity \ system answer	b	d	g	p	t	k
b	23	2				
d		16	1			
g			7			
p				9		1
t				1	18	1
k		1		4	5	11

Table 2: Stop identification matrix

ACKNOWLEDGMENTS

We would like to thank Caroline Huang, John Pitrelli, and Stephanie Seneff for serving as expert spectrogram readers.

REFERENCES

[1] R.A. Cole, A.I. Rudnicky, V.W. Zue, and D.R. Reddy, "Speech as Patterns on Paper," in *Perception and Production of Fluent Speech*, R.A. Cole, ed., Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980, pp. 3-50.

[2] R.A. Cole and V.W. Zue, "Speech as Eyes See It," in *Attention and Performance VIII*, R.S. Nickerson, ed. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980, pp. 475-494.

[3] V.W. Zue and R.A. Cole, "Experiments on Spectrogram Reading," *IEEE Conference Proceedings, ICASSP*, Washington D.C., 1979, pp. 116-119.

[4] J. Johannsen, J. MacAllister, T. Michalek, and S. Ross, "A Speech Spectrogram Expert," *IEEE Conference Proceedings, ICASSP*, Boston, MA, 1983, pp. 746-749.

[5] S.R. Johnson, J.H. Connolly, and E.A. Edmonds, "Spectrogram Analysis: A Knowledge-Based Approach to Automatic Speech Recognition," Leicester Polytechnic, Human Computer Interface Research Unit, Report No. 1, 1984.

[6] E.H. Shortliffe, *Computer Based Medical Consultations: MYCIN*, New York: American Elsevier Publishing Co., 1976.

[7] H.C. Leung and V.W. Zue, "Visual Characterization of Speech Spectrograms," *IEEE Conference Proceedings, ICASSP*, Tokyo, Japan, 1986, paper 51.1.

[8] H.C. Leung and V.W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech" *IEEE Conference Proceedings, ICASSP*, San Diego, CA, 1984, paper 2.9.

[9] N. Chomsky and M. Halle, *The Sound Pattern of English*, New York: Harper & Row, Pubs., 1968.

[10] P. Demichelis, R. De Mori, P. Laface, and M. O'Kane, "Computer Recognition of Plosive Sounds Using Contextual Information," *IEEE Trans. Acoust. Speech and Signal Process.*, Vol. ASSP-31, No. 2, Apr. 1983, pp. 359-377.

Paper presented at the 1985 IEEE-ASSPS Workshop on "Frontiers of Speech Recognition," Arden House, Harriman, NY, December 3-6, 1985.

Utilizing Speech-Specific Knowledge in Automatic Speech Recognition[1]

Victor W. Zue
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

In automatic speech recognition, the acoustic signal is the only tangible connection between the talker and the machine. While the signal conveys *linguistic* information, it also contains *extralinguistic* information about such matters as the identity of the speaker, his or her physiological and psychological states, and the acoustic environment. I believe that successful speech recognition is possible only if we can determine ways to extract the linguistic information while discarding irrelevant information.

Over the past three decades, we have made slow but steady progress in researching the complex relationship between the underlying linguistic representations of an utterance and its various acoustic realizations. While decades may pass before we reach a full understanding, we may still derive near-term benefits from the increased utilization of speech knowledge in speech recognition algorithms. The benefits can take the form of better algorithm performance or reduced sensitivity of systems to variations in speaker and environment.

In my presentation, I will suggest the following:

- Signal representation based on human auditory system may be important in enhancing phonetic contrasts.
- Performance of pattern recognition algorithms may be improved when augmented with speech knowledge.
- New models of speech recognition utilizing constraints imposed by the language may be effective.
- Optimum utilization of incomplete acoustic-phonetic knowledge in the form of ignorance modeling may be important.

---

[1]Research supported by DARPA contract N00014-82-K-0727, as monitored by the Office of Naval Research.

DISTRIBUTION LIST

	DODAAD Code	
Head Information Sciences Division Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217	N00014	(1)
Administrative Contracting Officer E19-628 Massachusetts Institute of Technology Cambridge, Massachusetts 02139		(1)
Director Naval Research Laboratory Washington, D.C. 20375 Attn: Code 2627	N00173	(1)
Defense Technical Information Center Bldg. 5, Cameron Station Alexandria, Virginia 22314	S47031	(12)

END

6-87

DTIC